EDS Activity 1

Name : Vedant Madankar

PRN : 202401040291

Roll No. : CS3-20

Dataset : Paper Review

Importing dataset into Jupyter Python Notebook

jupyter paper review Last Checkpoint: 6 minutes ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[1]:
```
import pandas as pd
df = pd.read_csv('reviews.csv')
df.head()
```

[1]:

	id	preliminary_decision	review__confidence	review__evaluation	review__id	review__lan	review__orientation	review__remarks	review__text	review__timespan
0	1.0	accept	4.0	1.0	1.0	es	0.0	NaN	- El artículo aborda un problema contingente y...	2010-07-05
1	NaN	NaN	4.0	1.0	2.0	es	1.0	NaN	El artículo presenta recomendaciones prácticas...	2010-07-05
2	NaN	NaN	5.0	1.0	3.0	es	1.0	NaN	- El tema es muy interesante y puede ser de mu...	2010-07-05
3	2.0	accept	4.0	2.0	1.0	es	1.0	NaN	Se explica en forma ordenada y didáctica una e...	2010-07-05
4	NaN	NaN	4.0	2.0	2.0	es	0.0	NaN	NaN	2010-07-05

[ ]:

1.Finding the number of missing values in each column.

jupyter paper review Last Checkpoint: 9 minutes ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

3	2.0	accept	4.0	2.0	1.0	es	1.0	NaN	forma ordenada y didáctica una e...	2010-07-05
4	NaN	NaN	4.0	2.0	2.0	es	0.0	NaN	NaN	2010-07-05

[2]:
```
import numpy as np

missing_values = df.isnull().sum()
print(missing_values)
```

```
id                     236
preliminary_decision   236
review__confidence       5
review__evaluation       3
review__id               3
review__lan              3
review__orientation      3
review__remarks        297
review__text             9
review__timespan         3
dtype: int64
```

[ ]:

2.Calculation of mean of the review confidence scores.

jupyter paper review Last Checkpoint: 10 minutes ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[4]:
```
mean_confidence = np.mean(df['review__confidence'].dropna())
print(mean_confidence)
```

```
3.5732009925558312
```

[ ]:

3.Find the most common review language and how many times it appears.

jupyter paper review Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[27]:
```
most_common_language = df['review__lan'].value_counts().idxmax()
most_common_count = df['review__lan'].value_counts().max()
print(most_common_language, most_common_count)
```

```
es 388
```

[ ]:

4.Filter reviews where review confidence is in the top 10% of all values.

jupyter paper review Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[28]:
```
threshold = np.percentile(df['review__confidence'].dropna(), 90)
top_10_confidence_reviews = df[df['review__confidence'] >= threshold]
print(top_10_confidence_reviews[['review__confidence', 'review__evaluation']])
```

```
     review__confidence  review__evaluation
2                   5.0                 1.0
13                  5.0                 2.0
17                  5.0                -2.0
21                  5.0                 1.0
22                  5.0                 2.0
28                  5.0                -1.0
41                  5.0                 2.0
71                  5.0                 2.0
75                  5.0                -2.0
85                  5.0                 2.0
117                 5.0                 1.0
118                 5.0                 0.0
133                 5.0                 1.0
134                 5.0                -2.0
136                 5.0                 2.0
137                 5.0                 1.0
146                 5.0                -2.0
148                 5.0                -2.0
152                 5.0                 2.0
155                 5.0                 2.0
164                 5.0                 2.0
166                 5.0                 2.0
167                 5.0                 2.0
172                 5.0                 1.0
176                 5.0                 2.0
```

5.Create a new column confidence category:

jupyter paper review Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[29]:
```
def categorize_confidence(x):
    if x <= 2:
        return 'Low'
    elif x <= 4:
        return 'Medium'
    else:
        return 'High'

df['confidence_category'] = df['review__confidence'].apply(categorize_confidence)
print(df[['review__confidence', 'confidence_category']].head())
```

```
   review__confidence confidence_category
0                 4.0             Medium
1                 4.0             Medium
2                 5.0               High
3                 4.0             Medium
4                 4.0             Medium
```

[ ]:

6.Find the average review evaluation for each confidence category.

jupyter paper review Last Checkpoint: 1 hour ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[30]:
```
avg_eval_by_conf_cat = df.groupby('confidence_category')['review__evaluation'].mean()
print(avg_eval_by_conf_cat)
```

```
confidence_category
High      0.000000
Low       0.121212
Medium    0.216718
Name: review__evaluation, dtype: float64
```

[ ]:

7.Find the 5 reviews with the highest confidence and their evaluation scores.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[31]:
```
top_5_confident_reviews = df.nlargest(5, 'review__confidence')[['review__confidence', 'review__evaluation']]
print(top_5_confident_reviews)
```

```
    review__confidence  review__evaluation
2                  5.0                 1.0
13                 5.0                 2.0
17                 5.0                -2.0
21                 5.0                 1.0
22                 5.0                 2.0
```

[ ]:

8.Calculate how many preliminary decisions are missing and what percentage that represents.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[32]:
```
missing_decisions = df['preliminary_decision'].isnull().sum()
percentage_missing_decisions = (missing_decisions / len(df)) * 100
print(f"Missing decisions: {missing_decisions} ({percentage_missing_decisions:.2f}%)")
```

```
Missing decisions: 236 (57.84%)
```

[ ]:

9.Create a summary dataframe with counts of 'Low', 'Medium', and 'High' confidence categories.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[33]:
```
confidence_summary = df['confidence_category'].value_counts()
print(confidence_summary)
```

```
confidence_category
Medium    323
High       52
Low        33
Name: count, dtype: int64
```

[ ]:

10.Identify columns where more than 30% of the data is missing. List those columns.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[35]:
```
missing_percentage = df.isnull().mean() * 100
columns_with_high_missing = missing_percentage[missing_percentage > 30].index.tolist()
print(columns_with_high_missing)
```

```
['id', 'preliminary_decision', 'review__remarks']
```

[ ]:

11.Calculate the interquartile range (IQR) of review evaluation and identify outlier reviews.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[36]:
```
Q1 = df['review__evaluation'].quantile(0.25)
Q3 = df['review__evaluation'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['review__evaluation'] < (Q1 - 1.5 * IQR)) | (df['review__evaluation'] > (Q3 + 1.5 * IQR))]
print(outliers[['review__evaluation']])
```

```
Empty DataFrame
Columns: [review__evaluation]
Index: []
```

[ ]:

12.Find the number of missing values in each column.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[38]:
```
print(df.isnull().sum())
```

```
id                      236
preliminary_decision    236
review__confidence        5
review__evaluation        3
review__id                3
review__lan               3
review__orientation       3
review__remarks         297
review__text              0
review__timespan          3
confidence_category       0
review_year               3
review_length             0
dtype: int64
```

[ ]:

13.Check if any duplicate review texts exist.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[40]:
```
duplicate_reviews = df.duplicated('review__text').sum()
print(duplicate_reviews)
```

```
8
```

[ ]:

14.Find the average length of review texts.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[41]:
```
df['review_length'] = df['review__text'].apply(lambda x: len(x))
average_length = df['review_length'].mean()
print(average_length)
```

```
998.0
```

[ ]:

15.Find the number of reviews written in the year 2015.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[42]:
```
reviews_2015 = df[df['review_year'] == 2015]
print(len(reviews_2015))
```

```
80
```

[ ]:

16.Get a summary of numeric columns using describe().

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[43]:
```
print(df.describe())
```

```
                id  review__confidence  review__evaluation  review__id  \
count  172.000000          403.000000          405.000000  405.000000
mean    86.500000            3.573201            0.182716    1.824691
std     49.796252            0.844341            1.502868    0.821362
min      1.000000            1.000000           -2.000000    1.000000
25%     43.750000            3.000000           -1.000000    1.000000
50%     86.500000            4.000000            0.000000    2.000000
75%    129.250000            4.000000            2.000000    2.000000
max    172.000000            5.000000            2.000000    4.000000

       review__orientation  review_year  review_length
count           405.000000   405.000000     408.000000
mean             -0.212346  2012.839506     998.000000
std               1.019292     1.919524     832.942786
min              -2.000000  2010.000000      18.000000
25%              -1.000000  2010.000000     462.750000
50%               0.000000  2014.000000     793.000000
75%               1.000000  2014.000000    1269.750000
max               2.000000  2015.000000    6345.000000
```

[ ]:

17.Find the earliest (oldest) review date.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[44]:
```
earliest_review = pd.to_datetime(df['review__timespan'].min()
print(earliest_review)
```

```
2010-07-05 00:00:00
```

[ ]:

18.Find how many reviews are written in English.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[45]:
```
english_reviews = (df['review__lan'] == 'English').sum()
print(english_reviews)
```

```
0
```

[ ]:

19.Find the maximum and minimum review confidence values.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[46]:
```
max_conf = df['review__confidence'].max()
min_conf = df['review__confidence'].min()
print(max_conf, min_conf)
```

```
5.0 1.0
```

[ ]:

20.Find the top 3 preliminary decisions having the highest average review evaluation.

jupyter paper review Last Checkpoint: 2 hours ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[47]:
```
top3_decisions = df.groupby('preliminary_decision')['review__evaluation'].mean().sort_values(ascending=False).head(3)
print(top3_decisions)
```

```
preliminary_decision
no decision        1.000000
accept             0.973451
probably reject   -1.142857
Name: review__evaluation, dtype: float64
```

[ ]: