EDS ASSIGNMENT

Name:- Siddhesh Santosh Shinde

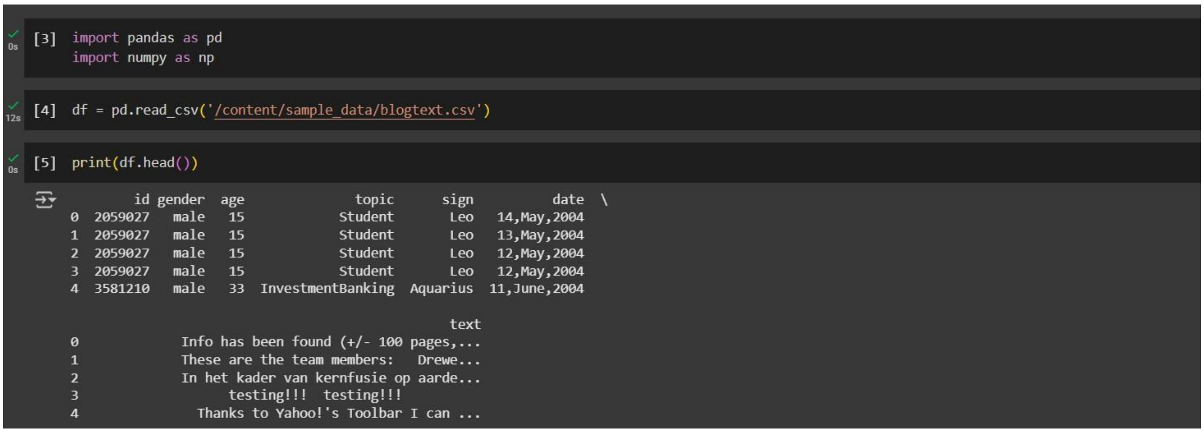
PRN:- 202401040176

Roll no.:- CS2-84

Batch:- C24

- Importing The Blog Authorship Corpus Dataset:-

1. Display first 5 rows of the dataset



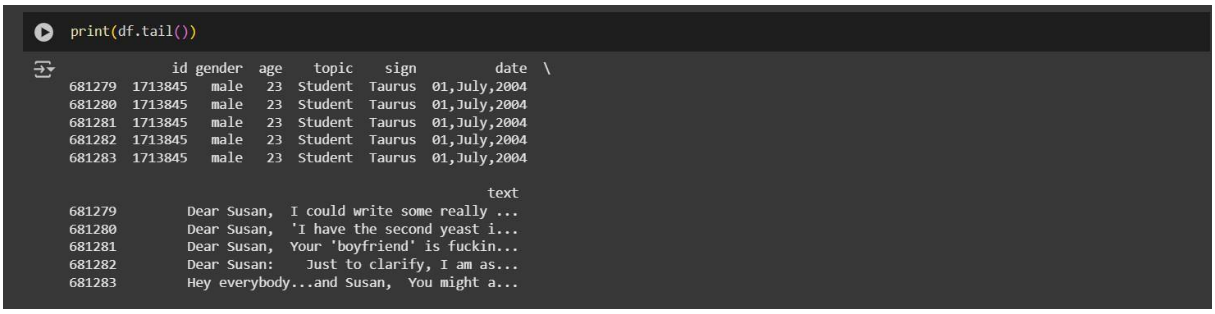
```
[3] import pandas as pd
    import numpy as np
```

```
[4] df = pd.read_csv('/content/sample_data/blogtext.csv')
```

```
[5] print(df.head())
```

```
        id gender  age              topic      sign          date  \
0  2059027   male   15            Student       Leo   14,May,2004   
1  2059027   male   15            Student       Leo   13,May,2004   
2  2059027   male   15            Student       Leo   12,May,2004   
3  2059027   male   15            Student       Leo   12,May,2004   
4  3581210   male   33  InvestmentBanking  Aquarius  11,June,2004   

                                             text  
0           Info has been found (+/- 100 pages,...  
1           These are the team members:   Drewe...  
2           In het kader van kernfusie op aarde...  
3                 testing!!!  testing!!!          
4             Thanks to Yahoo!'s Toolbar I can ...  
```

2. Display last 5 rows of the dataset

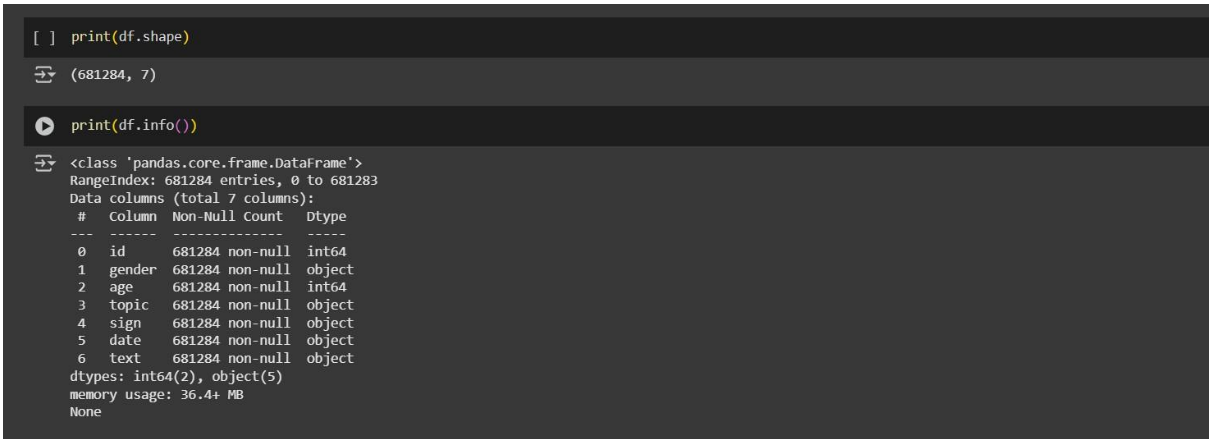


```
print(df.tail())
```

```
             id gender  age    topic    sign          date  \
681279  1713845   male   23  Student  Taurus  01,July,2004   
681280  1713845   male   23  Student  Taurus  01,July,2004   
681281  1713845   male   23  Student  Taurus  01,July,2004   
681282  1713845   male   23  Student  Taurus  01,July,2004   
681283  1713845   male   23  Student  Taurus  01,July,2004   

                                                     text  
681279         Dear Susan,  I could write some really ...  
681280         Dear Susan,  'I have the second yeast i...  
681281         Dear Susan,  Your 'boyfriend' is fuckin...  
681282         Dear Susan:    Just to clarify, I am as...  
681283         Hey everybody...and Susan,  You might a...  
```

3. Display shape of the dataset
4. Display information of the datset

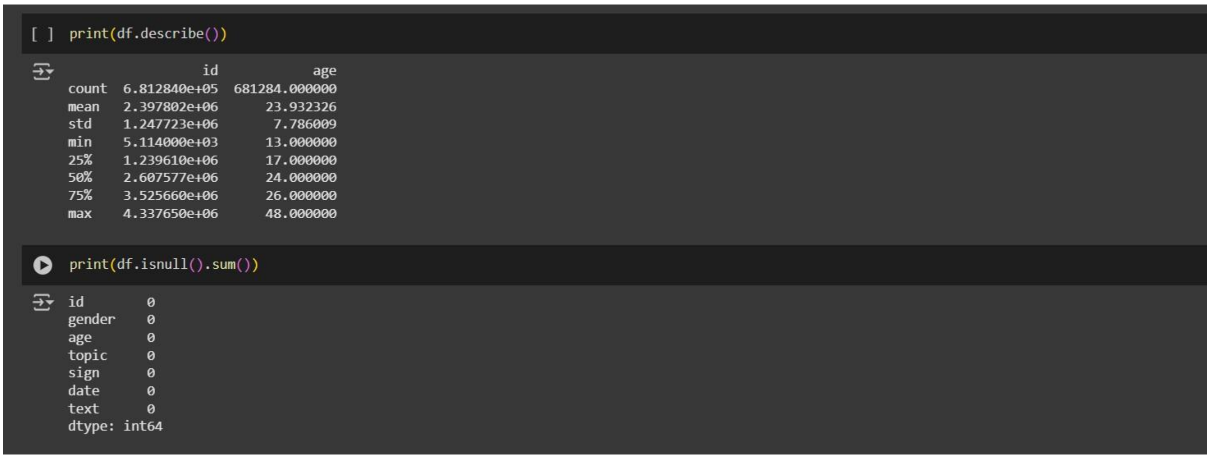
EDS ASSIGNMENT



```
[ ] print(df.shape)

(681284, 7)

print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 681284 entries, 0 to 681283
Data columns (total 7 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   id      681284 non-null  int64 
 1   gender  681284 non-null  object
 2   age     681284 non-null  int64 
 3   topic   681284 non-null  object
 4   sign    681284 non-null  object
 5   date    681284 non-null  object
 6   text    681284 non-null  object
dtypes: int64(2), object(5)
memory usage: 36.4+ MB
None
```

5. Describe the dataset

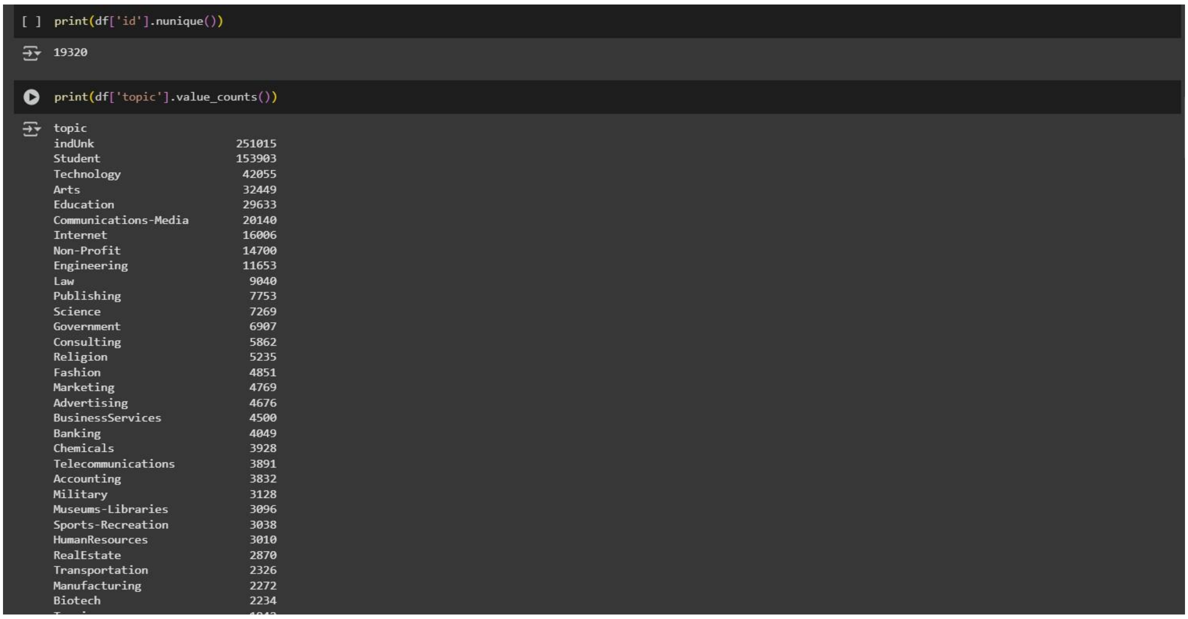
6. To find the missing values



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

                 id            age
count  6.812840e+05  681284.000000
mean   2.397802e+06      23.932326
std    1.247723e+06       7.786009
min    5.114000e+03      13.000000
25%    1.239610e+06      17.000000
50%    2.607577e+06      24.000000
75%    3.525660e+06      26.000000
max    4.337650e+06      48.000000

print(df.isnull().sum())

id        0
gender    0
age       0
topic     0
sign      0
date      0
text      0
dtype: int64
```

7. To check for duplicates

8. Value counts



```
[ ] print(df['id'].nunique())

19320

print(df['topic'].value_counts())

topic
indUnk                     251015
Student                    153903
Technology                  42055
Arts                        32449
Education                   29633
Communications-Media        20140
Internet                    16006
Non-Profit                  14700
Engineering                 11653
Law                          9040
Publishing                   7753
Science                      7269
Government                   6907
Consulting                   5862
Religion                     5235
Fashion                      4851
Marketing                    4769
Advertising                  4676
BusinessServices             4500
Banking                      4049
Chemicals                    3928
Telecommunications           3891
Accounting                   3832
Military                     3128
Museums-Libraries            3096
Sports-Recreation            3038
HumanResources               3010
RealEstate                   2870
Transportation               2326
Manufacturing                2272
Biotech                      2234
```

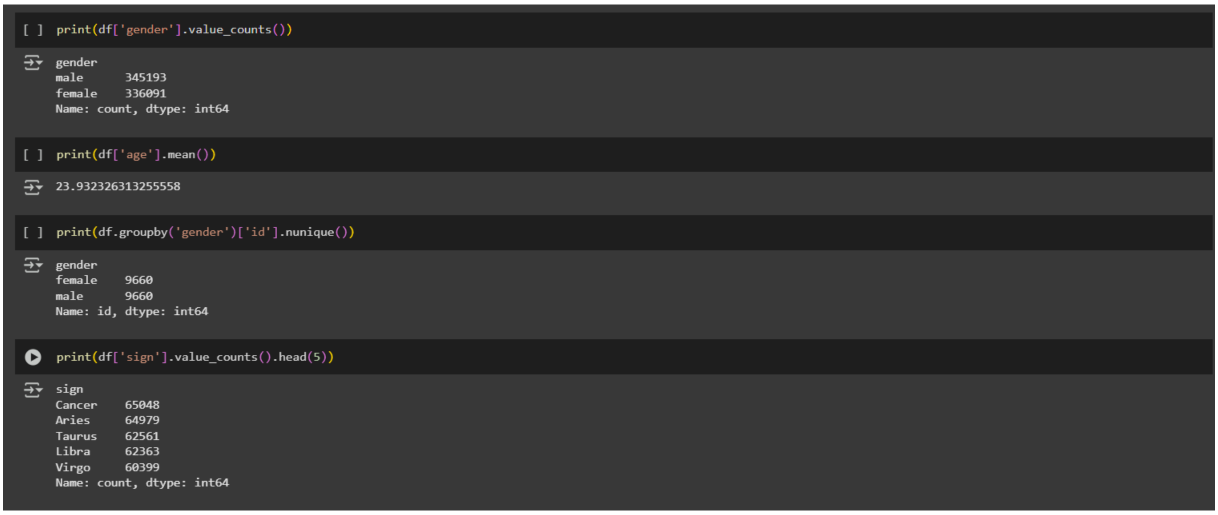
EDS ASSIGNMENT

9. Gender value count

10. Mean age

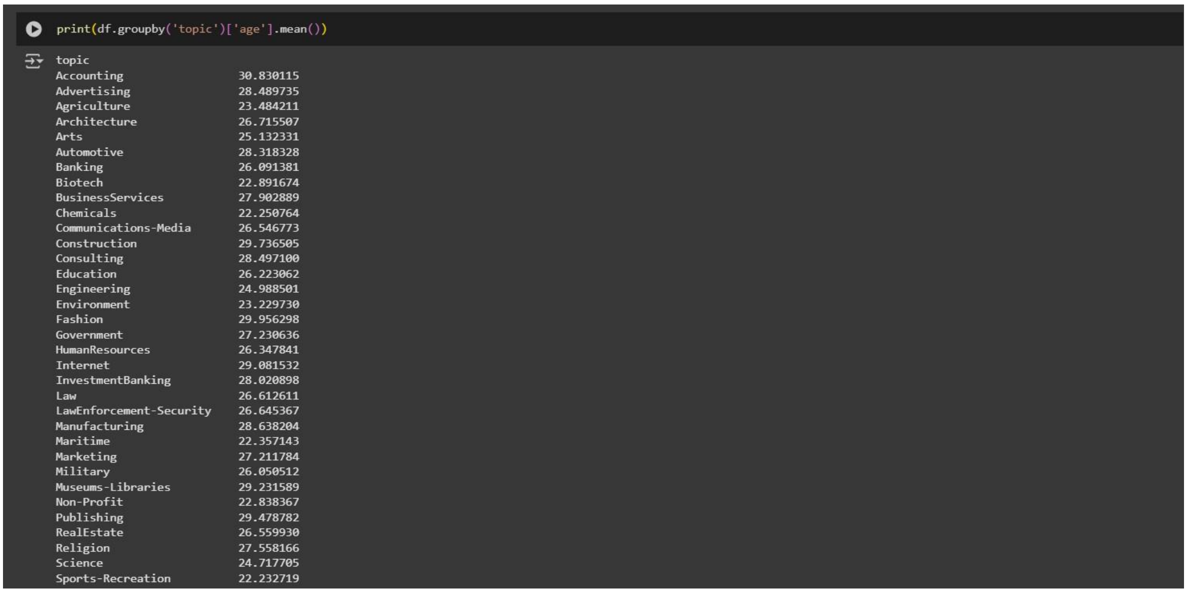
11. To check if the gender value count is duplicate

12. To see which zodiac signs are most common in your dataset



```
[ ] print(df['gender'].value_counts())

gender
male      345193
female    336091
Name: count, dtype: int64

[ ] print(df['age'].mean())

23.932326313255558

[ ] print(df.groupby('gender')['id'].nunique())

gender
female    9660
male      9660
Name: id, dtype: int64

print(df['sign'].value_counts().head(5))

sign
Cancer    65048
Aries     64979
Taurus    62561
Libra     62363
Virgo     60399
Name: count, dtype: int64
```

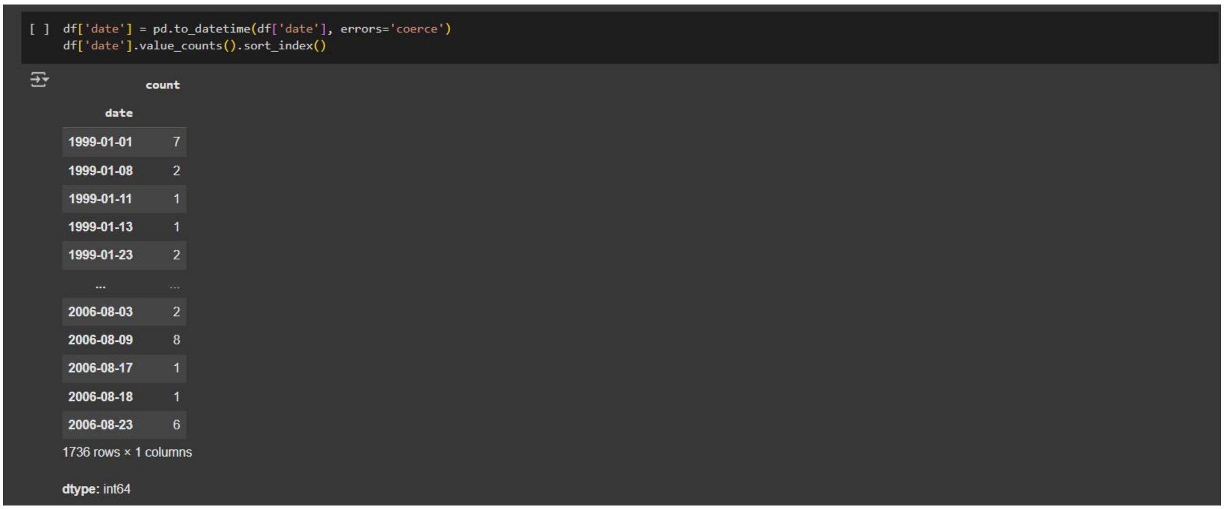
13. To summarize data quickly without manually filtering.



```
print(df.groupby('topic')['age'].mean())

topic
Accounting                 30.830115
Advertising                28.489735
Agriculture                23.484211
Architecture               26.715507
Arts                       25.132331
Automotive                 28.318328
Banking                    26.091381
Biotech                    22.891674
BusinessServices           27.902889
Chemicals                  22.250764
Communications-Media       26.546773
Construction               29.736505
Consulting                 28.497100
Education                  26.223062
Engineering                24.988501
Environment                23.229730
Fashion                    29.956298
Government                 27.230636
HumanResources             26.347841
Internet                   29.081532
InvestmentBanking          28.020898
Law                        26.612611
LawEnforcement-Security    26.645367
Manufacturing              28.638204
Maritime                   22.357143
Marketing                  27.211784
Military                   26.050512
Museums-Libraries          29.231589
Non-Profit                 22.838367
Publishing                 29.478782
RealEstate                 26.559930
Religion                   27.558166
Science                    24.717705
Sports-Recreation          22.232719
```

EDS ASSIGNMENT

14. To find how many blog posts (or records) happened on each date



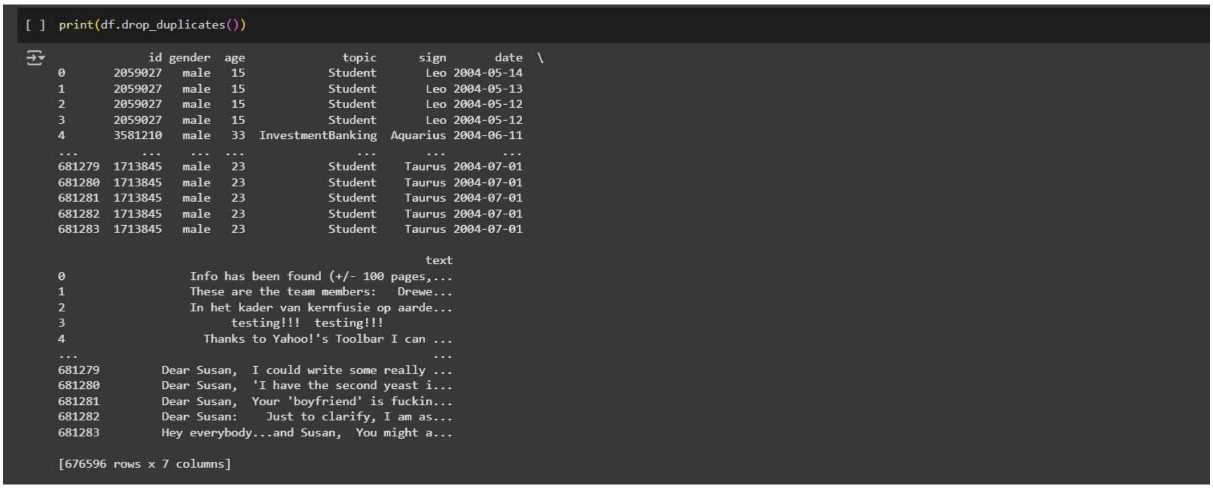
```
[ ] df['date'] = pd.to_datetime(df['date'], errors='coerce')
    df['date'].value_counts().sort_index()
```

	count
date	
1999-01-01	7
1999-01-08	2
1999-01-11	1
1999-01-13	1
1999-01-23	2
...	...
2006-08-03	2
2006-08-09	8
2006-08-17	1
2006-08-18	1
2006-08-23	6

1736 rows × 1 columns

dtype: int64

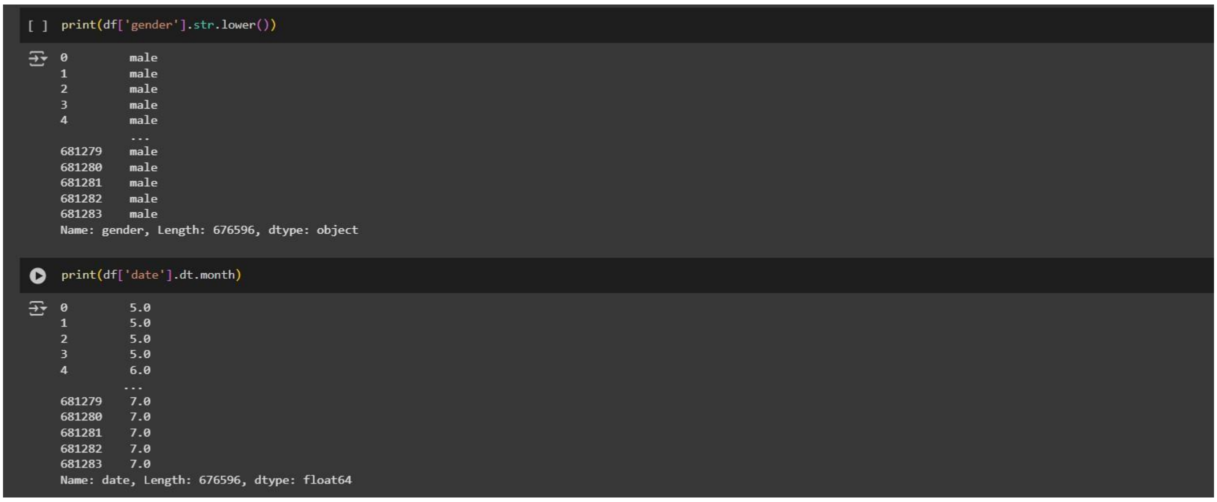
15. To find the duplicate values



```
[ ] print(df.drop_duplicates())
```

```
             id gender  age              topic      sign       date  \
0       2059027   male   15            Student       Leo 2004-05-14
1       2059027   male   15            Student       Leo 2004-05-13
2       2059027   male   15            Student       Leo 2004-05-12
3       2059027   male   15            Student       Leo 2004-05-12
4       3581210   male   33  InvestmentBanking  Aquarius 2004-06-11
...         ...    ...  ...                ...       ...        ...
681279  1713845   male   23            Student    Taurus 2004-07-01
681280  1713845   male   23            Student    Taurus 2004-07-01
681281  1713845   male   23            Student    Taurus 2004-07-01
681282  1713845   male   23            Student    Taurus 2004-07-01
681283  1713845   male   23            Student    Taurus 2004-07-01

                                                     text
0                   Info has been found (+/- 100 pages,...
1                   These are the team members:   Drewe...
2                   In het kader van kernfusie op aarde...
3                         testing!!!  testing!!!
4                     Thanks to Yahoo!'s Toolbar I can ...
...                                                   ...
681279         Dear Susan,  I could write some really ...
681280         Dear Susan,  'I have the second yeast i...
681281         Dear Susan,  Your 'boyfriend' is fuckin...
681282         Dear Susan:    Just to clarify, I am as...
681283         Hey everybody...and Susan,  You might a...

[676596 rows x 7 columns]
```

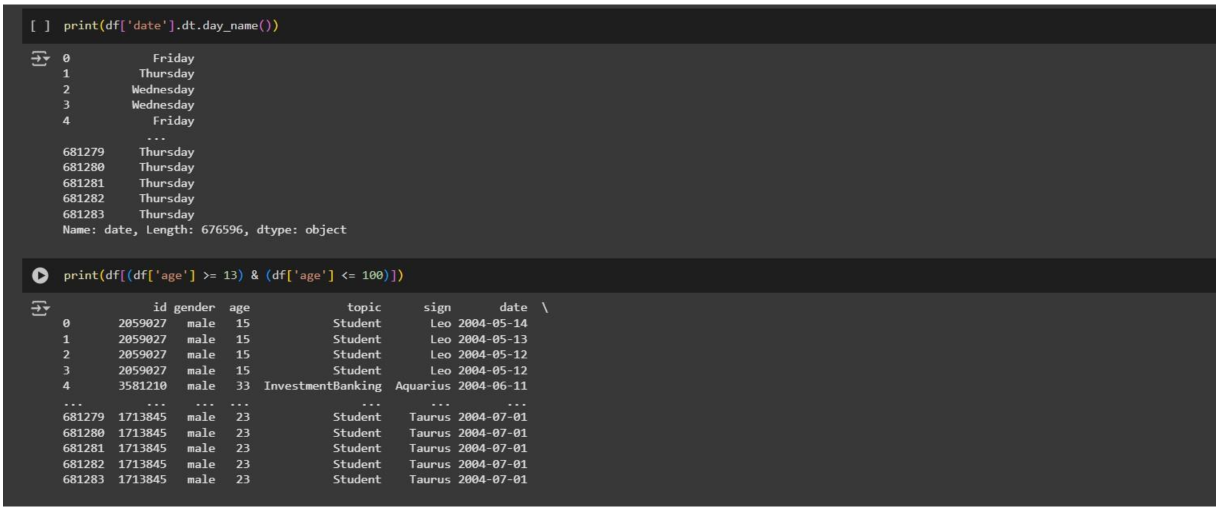
16. To standardize text
17. To pull out month information



```
print(df['gender'].str.lower())

0         male
1         male
2         male
3         male
4         male
          ...
681279    male
681280    male
681281    male
681282    male
681283    male
Name: gender, Length: 676596, dtype: object
```

```
print(df['date'].dt.month)

0         5.0
1         5.0
2         5.0
3         5.0
4         6.0
         ...
681279    7.0
681280    7.0
681281    7.0
681282    7.0
681283    7.0
Name: date, Length: 676596, dtype: float64
```

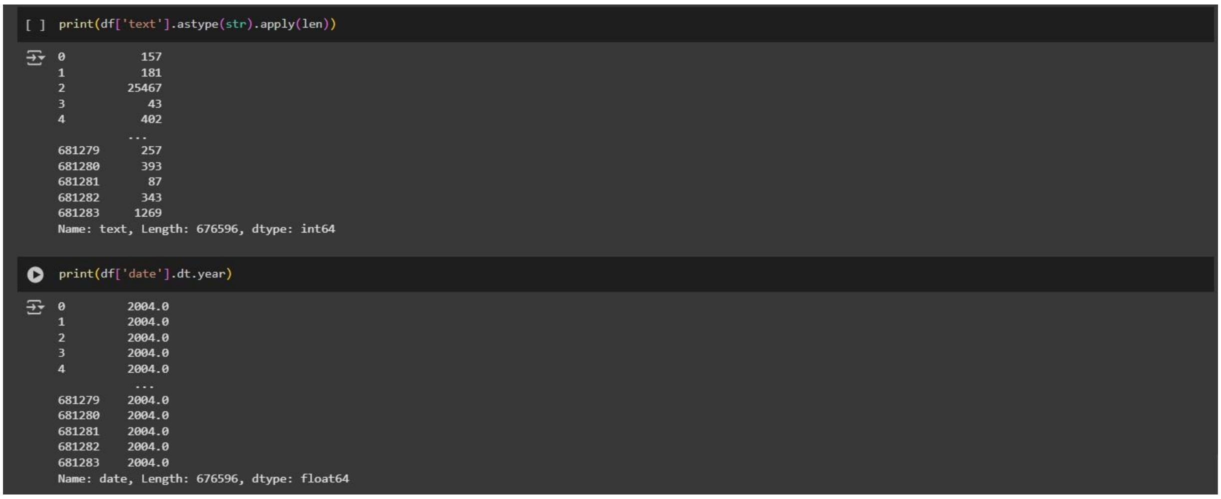
18. To know what day the records are made.
19. To focus only on sensible age groups.



```
print(df['date'].dt.day_name())

0            Friday
1          Thursday
2         Wednesday
3         Wednesday
4            Friday
             ...
681279     Thursday
681280     Thursday
681281     Thursday
681282     Thursday
681283     Thursday
Name: date, Length: 676596, dtype: object
```

```
print(df[(df['age'] >= 13) & (df['age'] <= 100)])

             id gender  age              topic      sign       date  \
0       2059027   male   15            Student       Leo 2004-05-14
1       2059027   male   15            Student       Leo 2004-05-13
2       2059027   male   15            Student       Leo 2004-05-12
3       2059027   male   15            Student       Leo 2004-05-12
4       3581210   male   33  InvestmentBanking  Aquarius 2004-06-11
...         ...    ...  ...                ...       ...        ...
681279  1713845   male   23            Student    Taurus 2004-07-01
681280  1713845   male   23            Student    Taurus 2004-07-01
681281  1713845   male   23            Student    Taurus 2004-07-01
681282  1713845   male   23            Student    Taurus 2004-07-01
681283  1713845   male   23            Student    Taurus 2004-07-01
```

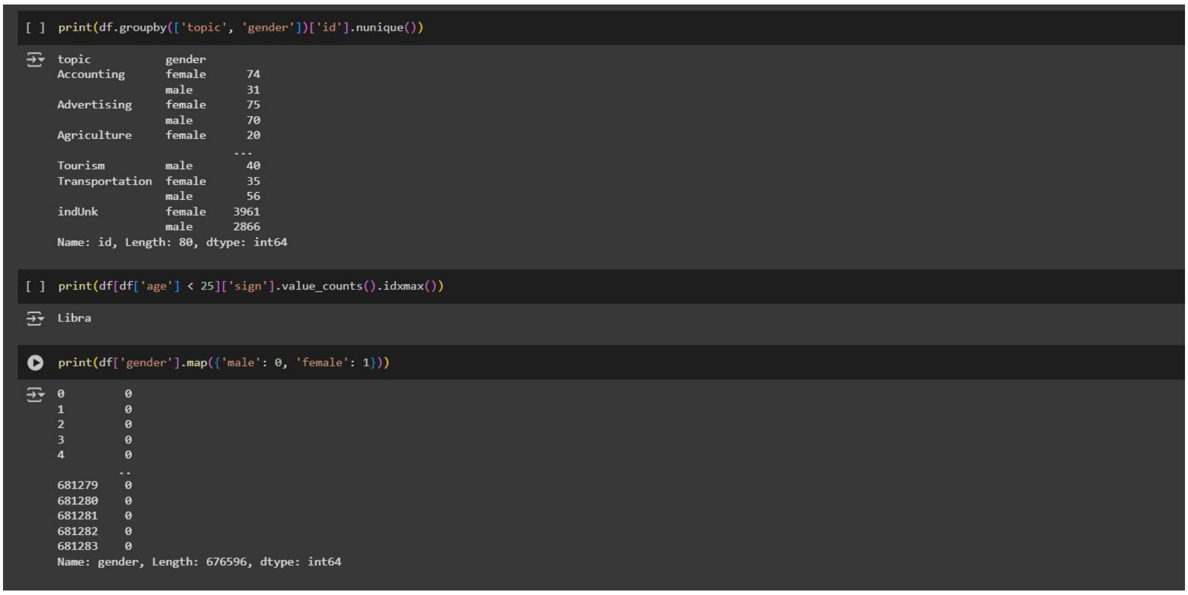
20. To calculate the Length of Text
21. To extract the Year from the Date



```
[ ] print(df['text'].astype(str).apply(len))

    0           157
    1           181
    2         25467
    3            43
    4           402
              ...
    681279      257
    681280      393
    681281       87
    681282      343
    681283     1269
    Name: text, Length: 676596, dtype: int64
```

```
print(df['date'].dt.year)

    0         2004.0
    1         2004.0
    2         2004.0
    3         2004.0
    4         2004.0
               ...
    681279    2004.0
    681280    2004.0
    681281    2004.0
    681282    2004.0
    681283    2004.0
    Name: date, Length: 676596, dtype: float64
```

22. To find unique IDs
23. To finds the most common zodiac sign under the age of 25
24. To convert gender into numeric form



```
[ ] print(df.groupby(['topic', 'gender'])['id'].nunique())

    topic           gender
    Accounting      female      74
                    male        31
    Advertising     female      75
                    male        70
    Agriculture     female      20
                               ...
    Tourism         male        40
    Transportation  female      35
                    male        56
    indUnk          female    3961
                    male      2866
    Name: id, Length: 80, dtype: int64
```

```
[ ] print(df[df['age'] < 25]['sign'].value_counts().idxmax())

    Libra
```

```
print(df['gender'].map({'male': 0, 'female': 1}))

    0         0
    1         0
    2         0
    3         0
    4         0
             ..
    681279    0
    681280    0
    681281    0
    681282    0
    681283    0
    Name: gender, Length: 676596, dtype: int64
```