accenture-vi

April 15, 2024

[1]:
```
import pandas as pd

# Read the three separate Excel files into Pandas DataFrames
content_df = pd.read_csv('/content/Content.csv')
reaction_df = pd.read_csv('/content/Reactions.csv')
reaction_types_df = pd.read_csv('/content/ReactionTypes.csv')
```

1 Data Cleaning

2 Content_df

[2]:
```
content_df.head()
```

[2]:
```
   Unnamed: 0                            Content ID  \
0           0  97522e57-d9ab-4bd6-97bf-c24d952602d2
1           1  9f737e0a-3cdd-4d29-9d24-753f4e3be810
2           2  230c4e4d-70c3-461d-b42c-ec09396efb3f
3           3  356fff80-da4d-4785-9f43-bc1261031dc6
4           4  01ab84dd-6364-4236-abbb-3f237db77180

                                User ID   Type        Category  \
0  8d3cd87d-8a31-4935-9a4f-b319bfe05f31  photo        Studying
1  beb1f34e-7870-46d6-9fc7-2e12eb83ce43  photo  healthy eating
2  a5c65404-5894-4b87-82f2-d787cbee86b4  photo  healthy eating
3  9fb4ce88-fac1-406c-8544-1a899cee7aaf  photo      technology
4  e206e31b-5f85-4964-b6ea-d7ee5324def1  video            food

                                                 URL
0  https://socialbuzz.cdn.com/content/storage/975…
1  https://socialbuzz.cdn.com/content/storage/9f7…
2  https://socialbuzz.cdn.com/content/storage/230…
3  https://socialbuzz.cdn.com/content/storage/356…
4  https://socialbuzz.cdn.com/content/storage/01a…
```

[3]:
```
# Check for missing values
print(content_df.isnull().sum())
```

```
Unnamed: 0      0
Content ID      0
User ID         0
Type            0
Category        0
URL           199
dtype: int64
```

[5]:

```python
# removing unnecessary columns
content_df.drop(['Unnamed: 0', 'User ID', 'URL'], axis =1, inplace= True)
```

[6]:

```python
content_df.isna().sum()
```

[6]:

```
Content ID    0
Type          0
Category      0
dtype: int64
```

[7]:

```python
content_df.groupby('Category').sum()
```

[7]:

```
                                                          Content ID  \
Category
"animals"                       5651450a-d330-46e6-bac9-c7c7e7defaf2
"cooking"                       b0782637-2604-40f7-8fc6-5e7d9ffd2255
"culture"          2d949603-6676-4402-900b-2c2c78315ea0693b5f91-5…
"dogs"             78461336-26e2-4d0c-ab9a-fefbe419a8dd9c8be342-6…
"food"                          e4487829-621f-4265-8665-42a4c0610745
"public speaking"               e1eb9e92-2e08-48b0-8be7-a2a98048aa08
"science"                       e4127cae-c2e9-4321-919a-be6d11636808
"soccer"           48a824bf-e495-4333-8acf-c800947ec2cd93081b1a-4…
"studying"                      c8b044a9-8427-4b41-bc61-7b549ab1626c
"technology"                    0fbdd670-a266-4805-9bbc-c5ad73cf97b3
"tennis"                        bde82663-7be3-453d-afc2-539202df27c3
"veganism"                      bc8024c3-c51e-4875-b8df-419e9848f492
Animals            07f88a73-aef2-45fd-8b5d-418e448b853d429632b9-8…
Culture                         6ddba21a-cc51-4bbd-b38d-02df5cec5f68
Education          45752c15-a54c-4b0d-8fe3-f39c40f6c8d92c043e74-6…
Fitness            409aa11d-5af3-4a70-9da1-482857f5835e3a106d1f-3…
Food               b2055111-9b7b-4a05-9f07-ac190a5391f0041bf6c9-8…
Healthy Eating                  279fdb2b-e9ca-4531-b55f-c8fe294083cd
Public Speaking                 37c5a8aa-d239-4b94-ae68-2754825d36ff
Science            6f48fe2b-7c20-4065-8ae9-7bea61275dc78b1bfacc-0…
Soccer             b27bfcfe-64df-499e-b60f-f93d2200c589e0c5ae74-b…
Studying           97522e57-d9ab-4bd6-97bf-c24d952602d2a372b4b7-6…
Technology                      ad5ddd13-b8ea-4174-ad71-da1663c7f959
Travel             2c0e1f1e-1af8-45b6-b8f2-9714358dd2ad410c757b-9…
Veganism                        62681a6c-fe82-4186-ba7d-0805ae5b95ed
```

```
animals            4fa14453-7b29-4302-b51f-9aa23b472c1b4478d98e-4…
cooking            cf1e8c1a-23eb-4426-9f58-002fb1b53e9102fa1c4f-7…
culture            259cd56f-b017-4a41-81a7-f26ce9b350925fbbdd47-e…
dogs               3f8590c7-6ab2-4973-805a-90cdec355f05809b41e3-7…
education          388bd9db-9d10-4f47-87c4-6db46e83bc95f08bdab2-b…
fitness            7ffd0a82-4a0a-4527-a4d6-e251b756bac7ab4c4756-1…
food               01ab84dd-6364-4236-abbb-3f237db7718081abd65a-3…
healthy eating     9f737e0a-3cdd-4d29-9d24-753f4e3be810230c4e4d-7…
public speaking    b18cb63f-4c8e-44ee-a47f-541e95191d1146fb701d-6…
science            5118e9c5-1377-4cc5-a486-65b35b7b7b7634a3747a-0…
soccer             0bedca96-fb76-4287-a83c-17330ed39ccef332d362-d…
studying           78d0075f-895c-4a15-a35c-a921e2bb2cea89fd8f89-8…
technology         356fff80-da4d-4785-9f43-bc1261031dc6e5490118-9…
tennis             0be59876-d70c-486c-8e0b-a06bef7a2cd6850fe90d-4…
travel             e6ee2244-9382-49a9-8cbf-fa54aaaa2392bda0b065-7…
veganism           2920dccb-e06f-49fc-8049-b6d4164dfe84bfa4e11c-9…

                                                                Type
Category
"animals"                                                      photo
"cooking"                                                      video
"culture"                                            audiophotoaudio
"dogs"                                                    videovideo
"food"                                                         audio
"public speaking"                                              video
"science"                                                      audio
"soccer"                                               audioGIFphoto
"studying"                                                     photo
"technology"                                                   photo
"tennis"                                                       video
"veganism"                                                       GIF
Animals                                             GIFaudiophotoGIF
Culture                                                        audio
Education                                                   photoGIF
Fitness                                      GIFphotovideovideoaudio
Food                                                          GIFGIF
Healthy Eating                                                 video
Public Speaking                                                photo
Science                                           GIFphotoaudiovideo
Soccer                                                 videoaudioGIF
Studying                                                    photoGIF
Technology                                                     video
Travel                                                    audiovideo
Veganism                                                       audio
animals            audioaudioGIFvideophotophotovideoGIFGIFGIFvide…
cooking            GIFvideoGIFGIFvideovideoGIFaudioaudioGIFvideov…
culture            videovideoaudiovideophotoGIFphotoaudioaudioaud…
```

```
dogs               videoaudiovideovideoaudiovideophotophotovideov…
education          videoaudioGIFphotovideoGIFphotovideophotoaudio…
fitness            GIFaudiophotoaudioGIFGIFphotoGIFaudioaudioGIFp…
food               videovideovideoGIFvideoaudioaudioGIFphotophoto…
healthy eating     photophotophotoGIFGIFGIFvideovideovideovideoph…
public speaking    photoaudiovideovideovideoGIFphotovideovideopho…
science            GIFGIFvideovideoaudiovideophotovideoGIFvideoph…
soccer             photoGIFGIFGIFGIFvideoGIFGIFvideoGIFaudioaudio…
studying           photoaudioGIFphotophotoGIFaudiophotoaudiovideo…
technology         photovideophotoGIFvideoGIFaudioaudioGIFaudioph…
tennis             GIFvideovideoGIFGIFaudioGIFvideoGIFvideovideov…
travel             audiophotoGIFGIFGIFphotophotoaudiovideovideoph…
veganism           GIFphotoGIFaudioGIFGIFvideovideovideophotoaudi…
```

[8]:
```
# replace some values

content_df['Category'] = content_df['Category'].str.replace('"','')
```

[9]:
```
content_df.groupby('Category').sum()
```

[9]:
```
                                                        Content ID  \
Category
Animals          07f88a73-aef2-45fd-8b5d-418e448b853d429632b9-8…
Culture                       6ddba21a-cc51-4bbd-b38d-02df5cec5f68
Education        45752c15-a54c-4b0d-8fe3-f39c40f6c8d92c043e74-6…
Fitness          409aa11d-5af3-4a70-9da1-482857f5835e3a106d1f-3…
Food             b2055111-9b7b-4a05-9f07-ac190a5391f0041bf6c9-8…
Healthy Eating                279fdb2b-e9ca-4531-b55f-c8fe294083cd
Public Speaking               37c5a8aa-d239-4b94-ae68-2754825d36ff
Science          6f48fe2b-7c20-4065-8ae9-7bea61275dc78b1bfacc-0…
Soccer           b27bfcfe-64df-499e-b60f-f93d2200c589e0c5ae74-b…
Studying         97522e57-d9ab-4bd6-97bf-c24d952602d2a372b4b7-6…
Technology                    ad5ddd13-b8ea-4174-ad71-da1663c7f959
Travel           2c0e1f1e-1af8-45b6-b8f2-9714358dd2ad410c757b-9…
Veganism                      62681a6c-fe82-4186-ba7d-0805ae5b95ed
animals          4fa14453-7b29-4302-b51f-9aa23b472c1b4478d98e-4…
cooking          cf1e8c1a-23eb-4426-9f58-002fb1b53e9102fa1c4f-7…
culture          259cd56f-b017-4a41-81a7-f26ce9b350922d949603-6…
dogs             3f8590c7-6ab2-4973-805a-90cdec355f05809b41e3-7…
education        388bd9db-9d10-4f47-87c4-6db46e83bc95f08bdab2-b…
fitness          7ffd0a82-4a0a-4527-a4d6-e251b756bac7ab4c4756-1…
food             01ab84dd-6364-4236-abbb-3f237db7718081abd65a-3…
healthy eating   9f737e0a-3cdd-4d29-9d24-753f4e3be810230c4e4d-7…
public speaking  b18cb63f-4c8e-44ee-a47f-541e95191d1146fb701d-6…
science          5118e9c5-1377-4cc5-a486-65b35b7b7b7634a3747a-0…
soccer           0bedca96-fb76-4287-a83c-17330ed39ccef332d362-d…
studying         78d0075f-895c-4a15-a35c-a921e2bb2cea89fd8f89-8…
```

```
technology       356fff80-da4d-4785-9f43-bc1261031dc6e5490118-9…
tennis           0be59876-d70c-486c-8e0b-a06bef7a2cd6850fe90d-4…
travel           e6ee2244-9382-49a9-8cbf-fa54aaaa2392bda0b065-7…
veganism         2920dccb-e06f-49fc-8049-b6d4164dfe84bfa4e11c-9…

                                                              Type
Category
Animals                                           GIFaudiophotoGIF
Culture                                                      audio
Education                                                 photoGIF
Fitness                                    GIFphotovideovideoaudio
Food                                                        GIFGIF
Healthy Eating                                               video
Public Speaking                                              photo
Science                                         GIFphotoaudiovideo
Soccer                                               videoaudioGIF
Studying                                                  photoGIF
Technology                                                   video
Travel                                                  audiovideo
Veganism                                                     audio
animals          audioaudioGIFvideophotophotovideophotoGIFGIFGI…
cooking          GIFvideoGIFGIFvideovideoGIFaudioaudioGIFvideov…
culture          videoaudiovideoaudiovideophotoGIFphotoaudioaud…
dogs             videoaudiovideovideoaudiovideophotophotovideov…
education        videoaudioGIFphotovideoGIFphotovideophotoaudio…
fitness          GIFaudiophotoaudioGIFGIFphotoGIFaudioaudioGIFp…
food             videovideovideoGIFvideoaudioaudioGIFphotophoto…
healthy eating   photophotophotoGIFGIFGIFvideovideovideovideoph…
public speaking  photoaudiovideovideovideoGIFphotovideovideopho…
science          GIFGIFvideovideoaudiovideophotovideoGIFvideoph…
soccer           photoGIFGIFGIFGIFvideoGIFGIFvideoGIFaudioaudio…
studying         photoaudioGIFphotophotoGIFaudiophotophotoaudio…
technology       photovideophotoGIFvideoGIFaudioaudioGIFaudioph…
tennis           GIFvideovideoGIFGIFaudioGIFvideoGIFvideovideov…
travel           audiophotoGIFGIFGIFphotophotoaudiovideovideoph…
veganism         GIFphotoGIFaudioGIFGIFvideovideovideophotoaudi…
```

[10]:
```
# Finding and remove duplicated values
```

[11]:
```
content_df.duplicated().sum()
```

[11]:
```
0
```

[12]:
```
content_df.drop_duplicates(subset=['Content ID'])
```

[12]:
```
                               Content ID   Type         Category
0    97522e57-d9ab-4bd6-97bf-c24d952602d2  photo         Studying
```

```
1    9f737e0a-3cdd-4d29-9d24-753f4e3be810  photo   healthy eating
2    230c4e4d-70c3-461d-b42c-ec09396efb3f  photo   healthy eating
3    356fff80-da4d-4785-9f43-bc1261031dc6  photo       technology
4    01ab84dd-6364-4236-abbb-3f237db77180  video             food
..                                    …      …                …
995  b4cef9ef-627b-41d7-a051-5961b0204ebb  video  public speaking
996  7a79f4e4-3b7d-44dc-bdef-bc990740252c    GIF       technology
997  435007a5-6261-4d8b-b0a4-55fdc189754b  audio         veganism
998  4e4c9690-c013-4ee7-9e66-943d8cbd27b7    GIF          culture
999  75d6b589-7fae-4a6d-b0d0-752845150e56  audio       technology

[1000 rows x 3 columns]
```

[13]:
```
content_df = content_df.drop_duplicates(subset=['Content ID'])
```

[14]:
```
# Standardize text data
content_df['Type'] = content_df['Type'].str.lower()
content_df['Category'] = content_df['Category'].str.lower()
```

[15]:
```
# Rename column name

content_df.rename(columns= {'Type': 'Content_Type'}, inplace= True)

content_df.columns
```

[15]:
```
Index(['Content ID', 'Content_Type', 'Category'], dtype='object')
```

3 Reaction_df

[16]:
```
reaction_df.head()
```

[16]:
```
   Unnamed: 0                            Content ID  \
0           0  97522e57-d9ab-4bd6-97bf-c24d952602d2
1           1  97522e57-d9ab-4bd6-97bf-c24d952602d2
2           2  97522e57-d9ab-4bd6-97bf-c24d952602d2
3           3  97522e57-d9ab-4bd6-97bf-c24d952602d2
4           4  97522e57-d9ab-4bd6-97bf-c24d952602d2

                                User ID     Type             Datetime
0                                   NaN      NaN  2021-04-22 15:17:15
1  5d454588-283d-459d-915d-c48a2cb4c27f  disgust  2020-11-07 09:43:50
2  92b87fa5-f271-43e0-af66-84fac21052e6  dislike  2021-06-17 12:22:51
3  163daa38-8b77-48c9-9af6-37a6c1447ac2   scared  2021-04-18 05:13:58
4  34e8add9-0206-47fd-a501-037b994650a2  disgust  2021-01-06 19:13:01
```

[17]:

```
# rename columns name

reaction_df.rename(columns= {'Type': 'Reaction_Type'}, inplace= True)
```

[18]:

```
reaction_df.isna().sum()
```

[18]:

```
Unnamed: 0          0
Content ID          0
User ID          3019
Reaction_Type     980
Datetime            0
dtype: int64
```

[19]:

```
# Since we are trying to find the top 5 categories by score.
# We removing user id, type which has null values
```

[20]:

```
reaction_df.drop(['User ID', 'Unnamed: 0'], axis = 1, inplace = True)
reaction_df.dropna(inplace=  True)
```

[21]:

```
reaction_df.isna().sum()
```

[21]:

```
Content ID       0
Reaction_Type    0
Datetime         0
dtype: int64
```

[22]:

```
# Remove duplicates
reaction_df = reaction_df.drop_duplicates(subset=['Content ID'])
```

[22]:

```

```

[23]:

```
# Format datetime
# Assuming Datetime column needs formatting
reaction_df['Datetime'] = pd.to_datetime(reaction_df['Datetime'],␣
 ↪format='%Y-%m-%d %H:%M:%S')


reaction_df.head()
```

```
<ipython-input-23-c4f57c64fac7>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  reaction_df['Datetime'] = pd.to_datetime(reaction_df['Datetime'],
format='%Y-%m-%d %H:%M:%S')
```

```
[23]:                                    Content ID Reaction_Type            Datetime
      1    97522e57-d9ab-4bd6-97bf-c24d952602d2       disgust 2020-11-07 09:43:50
      47   9f737e0a-3cdd-4d29-9d24-753f4e3be810       dislike 2020-06-25 17:01:58
      63   230c4e4d-70c3-461d-b42c-ec09396efb3f          hate 2020-12-07 19:13:23
      95   356fff80-da4d-4785-9f43-bc1261031dc6       worried 2021-04-03 12:17:12
      103  01ab84dd-6364-4236-abbb-3f237db77180        scared 2021-05-24 18:03:26
```

```
[23]:
```

```
[23]:
```

4 Reaction_types_df

```
[24]: reaction_types_df.shape
```

```
[24]: (16, 4)
```

```
[25]: reaction_types_df.head(20)
```

```
[25]:     Unnamed: 0         Type Sentiment  Score
      0            0        heart  positive     60
      1            1         want  positive     70
      2            2      disgust  negative      0
      3            3         hate  negative      5
      4            4   interested  positive     30
      5            5  indifferent   neutral     20
      6            6         love  positive     65
      7            7   super love  positive     75
      8            8      cherish  positive     70
      9            9        adore  positive     72
      10          10         like  positive     50
      11          11      dislike  negative     10
      12          12    intrigued  positive     45
      13          13      peeking   neutral     35
      14          14       scared  negative     15
      15          15      worried  negative     12
```

```
[26]: # rename columns name

      reaction_types_df.rename(columns= {'Type': 'Reaction_Type'}, inplace= True)
```

```
[27]: reaction_types_df.isna().sum()
```

```
[27]: Unnamed: 0       0
      Reaction_Type    0
      Sentiment        0
```

```
Score           0
dtype: int64
```

```
[28]: reaction_types_df.drop(['Unnamed: 0'], axis = 1, inplace = True)
```

```
[29]: reaction_types_df.drop_duplicates(subset= ['Reaction_Type'])
```

```
[29]:    Reaction_Type Sentiment  Score
      0          heart  positive     60
      1           want  positive     70
      2        disgust  negative      0
      3           hate  negative      5
      4     interested  positive     30
      5    indifferent   neutral     20
      6           love  positive     65
      7     super love  positive     75
      8        cherish  positive     70
      9          adore  positive     72
      10          like  positive     50
      11       dislike  negative     10
      12     intrigued  positive     45
      13       peeking   neutral     35
      14        scared  negative     15
      15       worried  negative     12
```

```
[29]:
```

5 Data cleaning completed

```
[30]: print(content_df.columns)
      print('\n')
      print(reaction_types_df.columns)
      print('\n')
      print(reaction_df.columns)
      print('\n')
```

```
Index(['Content ID', 'Content_Type', 'Category'], dtype='object')


Index(['Reaction_Type', 'Sentiment', 'Score'], dtype='object')


Index(['Content ID', 'Reaction_Type', 'Datetime'], dtype='object')
```

[33]:

```
# now trying to merge the cleaned datasets

# Merge the DataFrames
merged_df = pd.merge(reaction_df, content_df, on='Content ID', how='left')
merged_df = pd.merge(merged_df, reaction_types_df, on='Reaction_Type',␣
  ↪how='left')
merged_df.head()
```

[33]:

```
                             Content ID Reaction_Type            Datetime  \
0  97522e57-d9ab-4bd6-97bf-c24d952602d2       disgust 2020-11-07 09:43:50
1  9f737e0a-3cdd-4d29-9d24-753f4e3be810       dislike 2020-06-25 17:01:58
2  230c4e4d-70c3-461d-b42c-ec09396efb3f          hate 2020-12-07 19:13:23
3  356fff80-da4d-4785-9f43-bc1261031dc6       worried 2021-04-03 12:17:12
4  01ab84dd-6364-4236-abbb-3f237db77180        scared 2021-05-24 18:03:26

  Content_Type        Category Sentiment  Score
0        photo        studying  negative      0
1        photo  healthy eating  negative     10
2        photo  healthy eating  negative      5
3        photo      technology  negative     12
4        video            food  negative     15
```

[32]:

```
merged_df.columns
```

[32]:

```
Index(['Content ID', 'Reaction_Type', 'Datetime', 'Content_Type', 'Category'],
dtype='object')
```

[33]:

```

```

6 Finding Top 5 Categories with Highest Scores

[ ]:

```

```

[36]:

```
# Calculate the total scores for each category
category_scores = merged_df.groupby('Category')['Score'].sum().reset_index()

category_scores
```

[36]:

```
           Category  Score
0           animals   2299
1           cooking   2214
2           culture   2822
3              dogs   2375
4         education   2195
5           fitness   2500
6              food   2416
```

```
7    healthy eating   2457
8   public speaking   2083
9           science   2603
10           soccer   2281
11         studying   2133
12       technology   2567
13           tennis   1952
14           travel   2905
15         veganism   2482
```

7 Finding largest 5

[37]:

```
category_scores.nlargest(5, 'Score')
```

[37]:

```
      Category  Score
14      travel   2905
2      culture   2822
9      science   2603
12  technology   2567
5      fitness   2500
```

[38]:

```
# Sort the categories by score and get the top 5 performing categories
top_5_categories = category_scores.nlargest(5, 'Score')

# Display the top 5 performing categories
print(top_5_categories)
```

```
      Category  Score
14      travel   2905
2      culture   2822
9      science   2603
12  technology   2567
5      fitness   2500
```

[ ]:

```

```