Name: Farhan Shikalgar	Course Name: SMA Lab
Class: BECO	Roll no: 19CO56

Experiment No: 3

Aim: Data Cleaning and Storage- Preprocess, filter and store social media data for business (Using Python, MongoDB, R, etc).

Theory:

Social Media Scraping of Myntra using Python:

1. Instagram:

```
# Instagram
from instagramy import InstagramUser
user = InstagramUser("Myntra")
print(f"Username: {user.fullname}")
print(f"Biography: {user.biography}")
print(f"Verified User: {user.is_verified}")
print(f"Website: {user.website}")
print(f"Followers: {user.number_of_followers}")
print(f"Following: {user.number_of_followings}")
print(f"No. Of Posts: {user.number_of_posts}")
posts = user.posts
print(posts[0])
instaPosts = []
for i in range(10):
    post = {}
    post["Likes"] = posts[i].likes
    post["Comments"] = posts[i].comments
    post["post_source"] = posts[i].post_source
    post["post_url"] = posts[i].post_url
    post["time"] = posts[i].taken_at_timestamp
    instaPosts.append(post)
insta_df = pd.DataFrame(instaPosts)
insta_df.info()
```

```
25.print(insta_df.isna())
26.print(f"Latest posts:\n {insta_df}")
```

OUTPUT

```
└─ python Exp3-SMA.py
Username: MYNTRA
Biography: If you never try, you'll never know ❤
👨 #MyMyntraLook get featured on Myntra Studio
Visit Myntra Studio for more fashion & style advice.
Verified User: True
Website: https://myntra.onelink.me/eWcy/gnozbjm9
Followers: 3168336
Following: 262
No. Of Posts: 10804
Post(likes=283, comments=9, caption=None, is_video=True, timestamp=1675679094, location={'id': '109524955741121', 'has_public_page': True, 'name': 'India', 'slug': 'india'}, shortcode='CoUUcXvJzT6', post_url='https://www.instagram.com/p/CoUUcXvJzT6/', display_url='https://instagram.fbom56-1.fna.fbcdn.net/v/t51.2885-15/329543412_1342790036515421_8823910709516276393_n.jpg?stp=dst-jpg_e15&_nc_ht=instagram.fbom56-1.fna.fbcdn.net&_nc_cat=1&_nc_ohc=sEZNfq2LUdkAX_mCWQ_&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfAuruLbFXhitwlnp0VLfbppHaUgxzq70FfNuvlo9Up58w&oe=63E2D626&_nc_sid=7bff83', video_url='https://instagram.fbom56-1.fna.fbcdn.net/v/t66.30100-16/54642810_578501280529533_5147923975634763362_n.mp4?_nc_ht=instagram.fbom56-1.fna.fbcdn.net&_nc_cat=109&_nc_ohc=rKgHihplNFIAX_mxMkR&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfBJTQOv6StFY_MRmGFFz3XpgEmJhP9rYSz38cuKYINb-w&oe=63E274A7&_nc_sid=7bff83', video_view_count=4455, post_source='https://instagram.fbom56-1.fna.fbcdn.net/v/t66.30100-16/54642810_578501280529533_5147923975634763362_n.mp4?_nc_ht=instagram.fbom56-1.fna.fbcdn.net&_nc_cat=109&_nc_ohc=rKgHihplNFIAX_mxMkR&edm=ABfd0MgBAAAA&ccb=7-5&oh=00_AfBJTQOv6StFY_MRmGFFz3XpgEmJhP9rYSz38cuKYINb-w&oe=63E274A7&_nc_sid=7bff83', taken_at_timestamp=datetime.datetime(2023, 2, 6, 15, 54, 54))
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Likes        10 non-null     int64
 1   Comments     10 non-null     int64
```

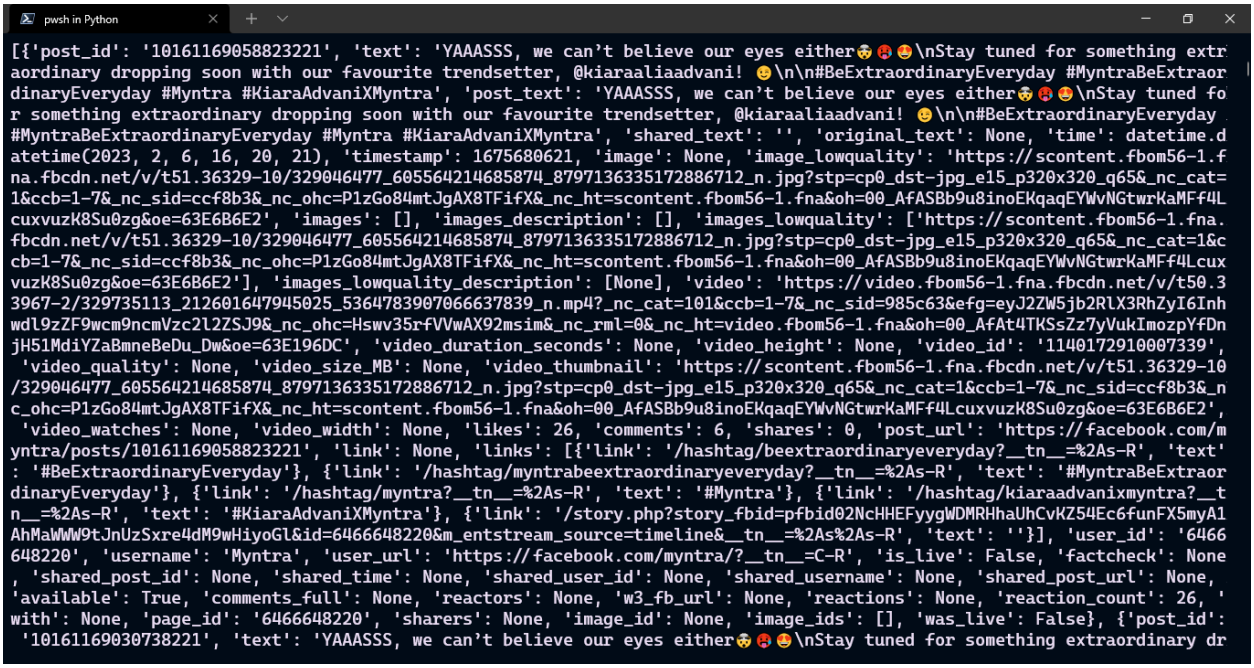
```
 4   time         10 non-null     datetime64[ns]
dtypes: datetime64[ns](1), int64(2), object(2)
memory usage: 384.0+ bytes
   Likes  Comments  post_source  post_url   time
0  False     False        False     False  False
1  False     False        False     False  False
2  False     False        False     False  False
3  False     False        False     False  False
4  False     False        False     False  False
5  False     False        False     False  False
6  False     False        False     False  False
7  False     False        False     False  False
8  False     False        False     False  False
9  False     False        False     False  False
Latest posts:
    Likes  Comments  ...                                  post_url               time
0     283         9  ...  https://www.instagram.com/p/CoUUcXvJzT6/ 2023-02-06 15:54:54
1     790        15  ...  https://www.instagram.com/p/CoTuf19D-xu/ 2023-02-06 10:27:51
2     327        10  ...  https://www.instagram.com/p/CoURNTXJ9zQ/ 2023-02-06 15:29:32
3     505        30  ...  https://www.instagram.com/p/CoUam3rP22C/ 2023-02-06 16:48:33
4     652         7  ...  https://www.instagram.com/p/CoUWZRxJCGx/ 2023-02-06 16:20:11
5    2156        12  ...  https://www.instagram.com/p/CoUVMPxs5la/ 2023-02-06 16:01:13
6     876         9  ...  https://www.instagram.com/p/CoSwXmrjct9/ 2023-02-06 01:23:10
7     796        21  ...  https://www.instagram.com/p/CoSEvupvhTz/ 2023-02-05 18:59:02
8    1320        55  ...  https://www.instagram.com/p/CoR3iSesDlg/ 2023-02-05 17:03:36
9    1092        23  ...  https://www.instagram.com/p/CoRj7jrpNPI/ 2023-02-05 14:12:17

[10 rows x 5 columns]
```

27.Facebook:

```
# Facebook
from facebook_scraper import get_posts
fbPosts = []
for post in get_posts('Myntra', pages=10):
    fbPosts.append(post)
print(fbPosts)
df = pd.DataFrame(fbPosts)
df.info()
df = df[['post_id', 'text', 'post_url', 'page_id', 'comments', 'likes',
   'time']]
print(df.isna())
print(print(f"Latest posts:\n {df}"))
```

OUTPUT



```
31    False  False     False    False     False  False  False
32    False  False     False    False     False  False  False
33    False  False     False    False     False  False  False
34    False  False     False    False     False  False  False
35    False  False     False    False     False  False  False
Latest posts:
            post_id                                          text   ...  likes               time
0   10161169058823221  YAAASSS, we can’t believe our eyes either😍😱🤩\n ...   ...   26.0 2023-02-06 16:20:21
1   10161169030738221  YAAASSS, we can’t believe our eyes either😍😱🤩\n ...   ...   10.0 2023-02-06 15:55:05
2   10161168987678221  YAAASSS, we can’t believe our eyes either😍😱🤩\n ...   ...    9.0 2023-02-06 15:29:42
3   10161168767013221  YAAASSS, we can’t believe our eyes either😍😱🤩\n ...   ...   11.0 2023-02-06 12:31:15
4   10161168690158221  YAAASSS, we can’t believe our eyes either😍😱🤩\n ...   ...   14.0 2023-02-06 11:15:04
5   10161167893718221  This Pink monochrome look is ❤️ ❤️\n\n📹 @simran ...   ...   29.0 2023-02-06 01:23:22
6   10161166649103221  Whether you’re going on a date, or giving your ...   ...   82.0 2023-02-05 14:12:22
7   10161165440628221  Mark your calendar & set your reminder 'cause  ...   ...   28.0 2023-02-04 23:38:35
8   10161162761178221  Jump into an unparalleled listening experience ...   ...   29.0 2023-02-04 19:46:55
9   10161164921723221  Shhhh…something extraordinary coming your way, ...   ...   70.0 2023-02-04 18:18:43
10  10161164580113221  Shhhh…something extraordinary coming your way, ...   ...   24.0 2023-02-04 13:36:57
11  10161164466943221  Shhhh…something extraordinary coming your way, ...   ...   20.0 2023-02-04 11:55:11
12  10161164422793221  Shhhh…something extraordinary coming your way, ...   ...   27.0 2023-02-04 11:07:25
13  10161163521578221  Denim game strong with pantaloons!\nCheck out  ...   ...   34.0 2023-02-03 23:55:34
14  10161163439843221  A race before the actual race!\nThe Early Bird ...   ...   29.0 2023-02-03 23:09:04
15  10161163209568221                                  Keep swiping 😉   ...  121.0 2023-02-03 21:12:09
16  10161161257373221  Vibe check 🖤\n🎥 @phoolfilledvish\n#Reels #Fash ...   ...   62.0 2023-02-03 00 :12:56
17  10161161191548221  💔\n🎥 @thankgod_itsfashion\n#Reels #FashionReel ...   ...   19.0 2023-02-02 23 :41:28
18  10161161063803221  The clock is ticking! ⏰\nYou don't wanna miss  ...   ...   45.0 2023-02-02 22:38:48
19  10161160787473221  Pastel magic!\nCheck out✨Myntra Studio ✨for fa ...   ...   60.0 2023-02-02 20:30:45
20  10161158857943221  Cute outfit inspo🤩\nCheck out✨Myntra Studio ✨f ...   ...    NaN 2023-02-02 00:01:45
21  10161158772288221  Tea party ready!🥰\nCheck out our stunning rang ...   ...   81.0 2023-02-01 23:23:33
```

39.Twitter:

```
# Twitter
import twint
c = twint.Config()
c.Lang = "en"
c.Username = "Myntra"
c.Pandas = True
# Run
twint.run.Search(c)
Tweets_df = twint.storage.panda.Tweets_df
Tweets_df.info()
print(Tweets_df.isna())
Tweets_df = Tweets_df[Tweets_df.language == "en"]
Tweets_df = Tweets_df[['id', 'date', 'nlikes', 'language', 'nreplies',
   'tweet', 'username']]
print(Tweets_df)
```

Output

```
1622578947249811456 2023-02-06 18:22:21 +0530 <myntra> @srikanth_vittam That doesn't sound like a happy situation, Vittam
! I've prioritized this concern &amp;amp; our case manager will contact you at the earliest. -KB  https://t.co/WsLuXFYsSb
1622578534584836096 2023-02-06 18:20:43 +0530 <myntra> @Mumalrajawat63 I get that you're expecting an update, and it was
never our objective to keep you waiting, Mumal! Upon checking, I see that an email was sent to you under reference number
 'IN23020316521082206453' on 06-Feb-2023 at 11:46 AM. Kindly check (cont)  https://t.co/CbJqJhBR2o
1622578068916404224 2023-02-06 18:18:52 +0530 <myntra> @bhavika_arora_ Bhavika! Upon checking, I see that one of our case
 managers tried calling you today, 12:40 PM, but unfortunately, could not connect. A notification was sent to you via the
 Myntra App with the reference number IN23020515312830651385. You (cont)  https://t.co/TjL3hTmwJq
1622578026235195394 2023-02-06 18:18:42 +0530 <myntra> @MuskanRahul6 Hi Upendra! I have replied to you here  https://t.co
/qRsSJRKsim . Please check. - ST
1622577891925200896 2023-02-06 18:18:10 +0530 <myntra> @MuskanRahul6 To hear this concerns us, Upendra! I regret the hass
le caused. Rest assured, we're closely following up with the logistics partners to have the product picked up from you so
on. One of our case managers will connect with you to share an update as early as possible. - ST
1622576957711056896 2023-02-06 18:14:27 +0530 <myntra> @singhsa1 Hi Sakshi! As we can see that one of our case managers c
ontacted you on 01 Feb 2023, 01:15 PM and ask for images of the product with bit more information. Also, send an in-app w
ith reference id IN23013111020246736311. Kindly have a check and revert for further assistance. -AB
1622576901993955328 2023-02-06 18:14:14 +0530 <myntra> @srikanth_vittam It was never our intention to cause any disruptio
n to our customers, Srikanth! Please be assured that I've highlighted your concern, and we are looking into it with the u
tmost priority. One of our case managers will connect with you at the earliest to provide an update. -MK
1622575586232377344 2023-02-06 18:09:00 +0530 <myntra> @Sumitzrd It's never our intention to keep you waiting for the res
olution, Sumit! Please be assured that we are working towards a resolution on priority, and one of our case managers will
 connect within the shortest time possible to address your concern with better precision. -MH
1622574933628051458 2023-02-06 18:06:24 +0530 <myntra> @MuskanRahul6 Upendra! Rest assured, I've prioritized this concern
 &amp;amp; one of our case managers will reach out to you at the earliest.- PK
1622574892456742912 2023-02-06 18:06:15 +0530 <myntra> @mahima_nayyar Hi Mahima! As we can see that one of our case manag
ers contacted you on 06 Feb 2023, 02:19 PM and ask for necessary information. Also, send an in-app with reference id IN23
020500492948796241. Kindly have a check and revert. Awaiting your response. -AB
1622574431993479172 2023-02-06 18:04:25 +0530 <myntra> @Rashi42691209 I hear you, Rashi! Upon checking, I see our case ma
nager has sent an In-App communication under the reference number IN23020312065441905756. You can track and view the same
```

```
pwsh in Python
Data columns (total 38 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               20 non-null     object
 1   conversation_id  20 non-null     object
 2   created_at       20 non-null     float64
 3   date             20 non-null     object
 4   timezone         20 non-null     object
 5   place            20 non-null     object
 6   tweet            20 non-null     object
 7   language         20 non-null     object
 8   hashtags         20 non-null     object
 9   cashtags         20 non-null     object
 10  user_id          20 non-null     int64
 11  user_id_str      20 non-null     object
 12  username         20 non-null     object
 13  name             20 non-null     object
 14  day              20 non-null     int64
 15  hour             20 non-null     object
 16  link             20 non-null     object
 17  urls             20 non-null     object
 18  photos           20 non-null     object
 19  video            20 non-null     int64
 20  thumbnail        20 non-null     object
 21  retweet          20 non-null     bool
 22  nlikes           20 non-null     int64
 23  nreplies         20 non-null     int64
 24  nretweets        20 non-null     int64
 25  quote_url        20 non-null     object
```

Conclusion:

We have successfully cleaned Myntra’s social media data from websites like Instagram, Facebook, and Twitter using various data preprocessing techniques.