

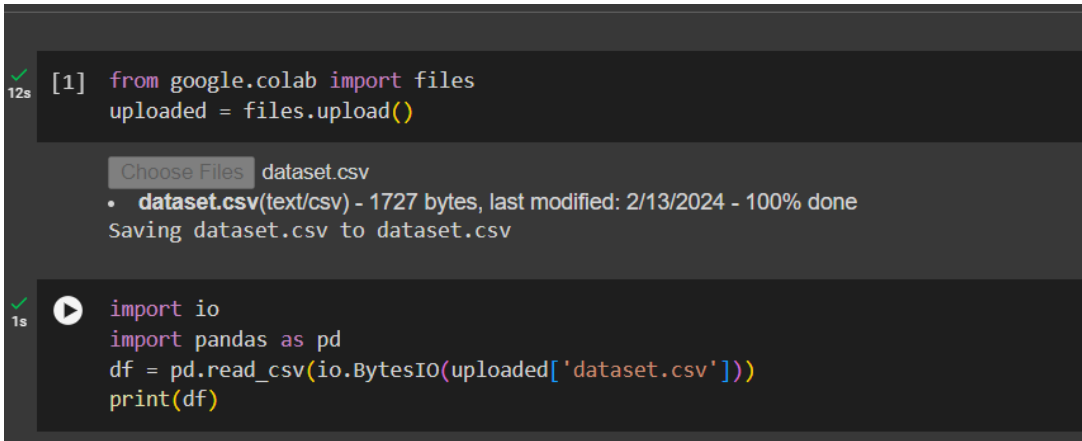
**Experiment No: 3**

Name: Alston Fernandes

Roll No.: 19

Batch: C

Performance Date: 14 February 2024

<b>Topic:</b>	Data Cleaning and Storage- Preprocess, filter and store social media data for business (Using Python, MongoDB, R, etc).
<b>Prerequisite:</b>	Knowledge of Social Media concepts, Data cleaning methods
<b>Mapping With COs:</b>	CSDL8023.3
<b>Objective:</b>	Students will be able to preprocess, filter and store social media data for analysis.
<b>Outcome:</b>	Students should be able to fix or remove incorrect, corrupted, incorrectly formatted, or duplicate, incomplete data within a dataset, which will increase overall productivity and allow for the highest quality information in their decision-making.
<b>Instructions:</b>	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
<b>Deliverables:</b>	<p><b>Colab Link:</b> <a href="https://colab.research.google.com/drive/1a8E47h1UIhgap-v_fs45tdSeAJ-2fn6L?usp=sharing">https://colab.research.google.com/drive/1a8E47h1UIhgap-v_fs45tdSeAJ-2fn6L?usp=sharing</a></p>  <p>The screenshot shows a Google Colab notebook. The first cell contains the code <code>from google.colab import files</code> and <code>uploaded = files.upload()</code>. Below the code, a file named <code>dataset.csv</code> is shown as uploaded, with a size of 1727 bytes and a status of '100% done'. The second cell contains the code <code>import io</code>, <code>import pandas as pd</code>, <code>df = pd.read_csv(io.BytesIO(uploaded['dataset.csv']))</code>, and <code>print(df)</code>.</p>

```
[2]      id      author      published_at      updated_at  \
0      0      @LeilaGharani  2024-01-25T17:16:47Z  2024-01-25T17:16:47Z
1      1      @truth5119    2023-12-29T07:35:07Z  2023-12-29T07:35:21Z
2      2      @michaeljarcher  2023-11-24T12:17:39Z  2023-11-24T12:17:39Z
3      3      @onewhitewolf13  2023-11-01T06:58:12Z  2023-11-01T06:58:12Z
4      4      @md.mahmudulalamsymon1216  2023-10-18T18:12:05Z  2023-10-18T18:12:05Z
5      5      @iditenahui    2023-09-23T18:50:10Z  2023-09-23T18:50:10Z
6      6      @siryoneyal    2023-09-06T09:13:18Z  2023-09-06T09:13:18Z
7      7      @kn4golf      2023-09-05T14:03:32Z  2023-09-05T14:03:32Z
8      8      @walterf6763    2023-09-04T21:57:03Z  2023-09-04T21:57:03Z
9      9      @FutureCommentary1  2023-08-31T05:58:46Z  2023-08-31T05:58:45Z
10     10     @m_stedt      2023-08-28T11:24:58Z  2023-08-28T11:24:57Z
11     11     @wr3661      2023-08-27T23:44:04Z  2023-08-27T23:44:03Z
12     12     @SilverVillacacan  2023-08-25T10:33:41Z  2023-08-25T10:33:41Z
13     13     @RayMedina    2023-08-22T00:45:58Z  2023-08-22T00:45:57Z
14     14     @gui.liraaa    2023-08-18T20:45:41Z  2023-08-18T20:45:41Z

      like_count      text
0      0      <a href="UCJtU0os_MwJa_Ewii-R3cJA/INVYZKnjF56F...
1      0      I materials had code if i want both displayed ...
2      0      About time excel got with the picture. But fir...
3      1      Was this a phased release? My images are showi...
4      0      Wow!!<br>Thanks a lot for sharing these nice t...
5      0      This is super helpful and useful as well! Than...
6      0      @LeilaGharani , I have kind of a challenge for...
7      0      As always, great presentation ?? Works for me ...
8      0      Thanks for the class. Can the same be done wi...
9      1      When will it be available to general 365 users?
10     0      Love that you can now insert pictures into cel...
11     0      Awesome ??. I have a question. Can I create a ...
12     0      can you make videos how to run macros in prote...
13     0      Hi Leila, love the show! Thank you for all the...
14     0      In my Excel, I can find the Data &gt; Data Typ...
```

```
✓ 0s print(df['author'])

0      @LeilaGharani
1      @truth5119
2      @michaeljarcher
3      @onewhitewolf13
4      @md.mahmudulalamsymon1216
5      @iditenahui
6      @siryoneyal
7      @kn4golf
8      @walterf6763
9      @FutureCommentary1
10     @m_stedt
11     @wr3661
12     @SilverVillacacan
13     @RayMedina
14     @gui.liraaa
Name: author, dtype: object
```

```
[5] print(df['text'])
```

```

0    <a href="UCJtU0os_MwJa_Ewii-R3cJA/INVYZKnjF56F...
1    I materials had code if i want both displayed ...
2    About time excel got with the picture. But fir...
3    Was this a phased release? My images are showi...
4    Wow!!<br>Thanks a lot for sharing these nice t...
5    This is super helpful and useful as well! Than...
6    @LeilaGharani , I have kind of a challenge for...
7    As always, great presentation ?? Works for me ...
8    Thanks for the class. Can the same be done wi...
9    When will it be available to general 365 users?
10   Love that you can now insert pictures into cel...
11   Awesome ??. I have a question. Can I create a ...
12   can you make videos how to run macros in prote...
13   Hi Leila, love the show! Thank you for all the...
14   In my Excel, I can find the Data &gt; Data Typ...
Name: text, dtype: object

```

```
[6] # Take a look at the first few rows
print(df.head())
```

	id	author	published_at	updated_at	\
0	0	@LeilaGharani	2024-01-25T17:16:47Z	2024-01-25T17:16:47Z	
1	1	@truth5119	2023-12-29T07:35:07Z	2023-12-29T07:35:21Z	
2	2	@michaeljarcher	2023-11-24T12:17:39Z	2023-11-24T12:17:39Z	
3	3	@onewhitewolf13	2023-11-01T06:58:12Z	2023-11-01T06:58:12Z	
4	4	@md.mahmudalamsymon1216	2023-10-18T18:12:05Z	2023-10-18T18:12:05Z	

	like_count	text
0	0	<a href="UCJtU0os_MwJa_Ewii-R3cJA/INVYZKnjF56F...
1	0	I materials had code if i want both displayed ...
2	0	About time excel got with the picture. But fir...
3	1	Was this a phased release? My images are showi...
4	0	Wow!! Thanks a lot for sharing these nice t...

```
[7] print(df.columns[1:4])
```

```
Index(['author', 'published_at', 'updated_at'], dtype='object')
```

```
[9] # select three rows and two columns
df.loc[1:3, ['author', 'text']]
```

	author	text
1	@truth5119	I materials had code if i want both displayed ...
2	@michaeljarcher	About time excel got with the picture. But fir...
3	@onewhitewolf13	Was this a phased release? My images are showi...

 `df[df.columns[1:4]]`





	author	published_at	updated_at
0	@LeilaGharani	2024-01-25T17:16:47Z	2024-01-25T17:16:47Z
1	@truth5119	2023-12-29T07:35:07Z	2023-12-29T07:35:21Z
2	@michaeljarcher	2023-11-24T12:17:39Z	2023-11-24T12:17:39Z
3	@onewhitewolf13	2023-11-01T06:58:12Z	2023-11-01T06:58:12Z
4	@md.mahmudulamsymon1216	2023-10-18T18:12:05Z	2023-10-18T18:12:05Z
5	@iditenahui	2023-09-23T18:50:10Z	2023-09-23T18:50:10Z
6	@siryoneyal	2023-09-06T09:13:18Z	2023-09-06T09:13:18Z
7	@kn4golf	2023-09-05T14:03:32Z	2023-09-05T14:03:32Z
8	@walterf6763	2023-09-04T21:57:03Z	2023-09-04T21:57:03Z
9	@FutureCommentary1	2023-08-31T05:58:46Z	2023-08-31T05:58:45Z
10	@m_stedt	2023-08-28T11:24:58Z	2023-08-28T11:24:57Z
11	@wr3661	2023-08-27T23:44:04Z	2023-08-27T23:44:03Z
12	@SilverVillacacan	2023-08-25T10:33:41Z	2023-08-25T10:33:41Z
13	@RayMedina	2023-08-22T00:45:58Z	2023-08-22T00:45:57Z
14	@gui.liraaa	2023-08-18T20:45:41Z	2023-08-18T20:45:41Z

No charts were generated by quickchart

```
[10] # .loc DataFrame method
# filtering rows and selecting columns by label
# format
# df.loc[rows, columns]
# row 1, all columns
df.loc[0, :]
```

```
id
author
published_at
updated_at
like_count
text
Name: 0, dtype: object
```

```
 # Remember that Python does not
# slice inclusive of the ending index.
# select all rows
# select first two column
df.iloc[:, 0:2]
```

 `df.iloc[:, 0:2]`



	id	author
0	0	@LeilaGharani
1	1	@truth5119
2	2	@michaeljarcher
3	3	@onewhitewolf13
4	4	@md.mahmudulalamsymon1216
5	5	@iditenahui
6	6	@siryoneyal
7	7	@kn4golf
8	8	@walterf6763
9	9	@FutureCommentary1
10	10	@m_stedt
11	11	@wr3661
12	12	@SilverVillacacan
13	13	@RayMedina
14	14	@gui.liraaa



```
[12] # iloc[row slicing, column slicing]  
df.iloc [0:2, 1:3]
```

	author	published_at
0	@LeilaGharani	2024-01-25T17:16:47Z
1	@truth5119	2023-12-29T07:35:07Z



	<pre>[15] # count all values in like_count column print(df['like_count'].value_counts())  0    13 1     2 Name: like_count, dtype: int64</pre> <pre>print(df['like_count'].sum())  2</pre> <pre>[16] # get about age print(df['published_at'].describe())  count                15 unique                15 top      2024-01-25T17:16:47Z freq                  1 Name: published_at, dtype: object</pre>
<b>Conclusion:</b>	Able to successfully preprocess, filter and store social media data for analysis.
<b>References:</b>	<a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a> <a href="https://elearn.dbit.in/mod/resource/view.php?id=8562">https://elearn.dbit.in/mod/resource/view.php?id=8562</a> <a href="https://www.python.org/">https://www.python.org/</a>

# Don Bosco Institute of Technology

## Department of Computer Engineering

### Assessment Rubric for Experiment No. 3

**Title of Experiment:** Data Cleaning from social media data

**Year and Semester:** 4th Year and VIII<sup>th</sup> Semester

Sr. No.	Criteria	1 Marks	2 Marks	3 Marks	4 Marks	5 Marks
1	Productivity	Not Satisfactory	Satisfactory	Good	Very Good	Excellent
2	Performance (Implementation)	Not Satisfactory	Satisfactory	Good	Very Good	Excellent
3	Viva	Satisfactory	Good	Very Good		
4	Submission on Time	Submitted after the given deadline	Submitted before the given deadline			