In [6]:

```
#!pip install sqlite3
```

Data profiling is the process of examining, analyzing, and creating useful summaries of data.

In [7]:

```
import pandas as pd
from pandasql import sqldf
```

In [8]:

```
data=pd.read_csv('/Users/pragatigupta/Documents/AI And ML/Linkedin Pos
```

In [57]:

```
df = sqldf("SELECT * FROM data");
df.head()
```

Out[57]:

	ID	Student ID	Gender	AGE	Score	CLASS
0	1	17975	F	15	6.7	y
1	2	34221	M	16	6.5	y
2	3	47975	F	17	5.5	y
3	4	87656	F	14	6.8	y
4	5	34223	M	15	7.1	y

In [60]:

```
# Find the data types of columns in the DataFrame
column_data_types = df.dtypes
# Print or display the data types
print(column_data_types)
```

```
ID                int64
Student ID       object
Gender           object
AGE               int64
Score           float64
CLASS            object
dtype: object
```

Get basic statistics

```
In [26]: # Total Count
         Total_IDs = sqldf("SELECT count() As Total_IDs From df");
         Total_IDs
```

Out[26]:

	Total_IDs
0	30

In [46]:

```
#Duplicates
Total_IDs = sqldf("SELECT count() As Total_IDs From df Group By ID ");
Total_IDs
```

Out[46]:

	Total_IDs
0	1
1	2
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1

In [40]:

```
Score_Avg = sqldf("SELECT Avg(Score) AS Score_Avg From df");
Score_Avg
```

Out[40]:

	Score_Avg
0	5.73

- Check for missing values

In [43]:

```
missing_count=sqldf("SELECT ID, COUNT(*) AS missing_count FROM df GROU
missing_count
```

Out[43]:

	ID	missing_count
0	2	2
1	30	1
2	29	1
3	28	1
4	27	1
5	26	1
6	25	1
7	24	1
8	23	1
9	22	1
10	21	1
11	20	1
12	19	1
13	18	1
14	17	1
15	16	1
16	15	1
17	14	1
18	12	1
19	11	1
20	10	1
21	9	1
22	8	1
23	7	1
24	6	1
25	5	1
26	4	1
27	3	1
28	1	1

-- Check for duplicate rows

In [44]:

```
SELECT ID, COUNT(*) AS duplicate_count FROM df GROUP BY ID HAVING COUNT
```

Out[44]:

	ID	duplicate_count
0	2	2

In [ ]: