

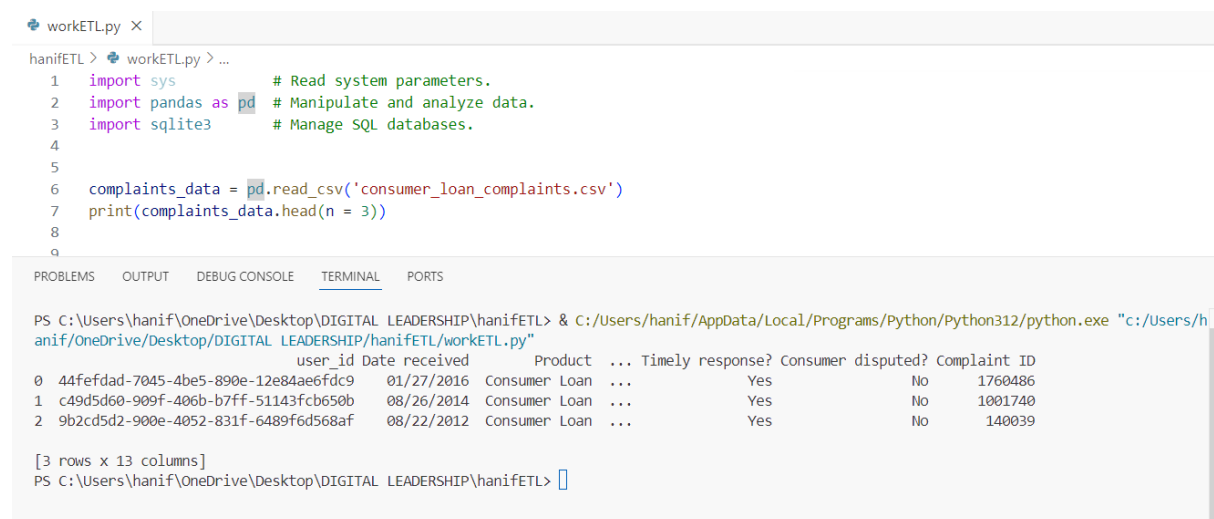
## ETL - Extracting, Transforming, and Loading Data

ETL: The process of combining data from multiple sources, preparing the data, and loading the result into a destination.

### Objective

- learn to extract data from various data
- learn to transform data into a proper format
- learn to load data into a destination

### 1)Extract Data



```
workETL.py X
hanifETL > workETL.py > ...
1 import sys # Read system parameters.
2 import pandas as pd # Manipulate and analyze data.
3 import sqlite3 # Manage SQL databases.
4
5
6 complaints_data = pd.read_csv('consumer_loan_complaints.csv')
7 print(complaints_data.head(n = 3))
8
9
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
```

	user_id	Date received	Product	...	Timely response?	Consumer disputed?	Complaint ID
0	44fefdad-7045-4be5-890e-12e84ae6fdc9	01/27/2016	Consumer Loan	...	Yes	No	1760486
1	c49d5d60-909f-406b-b7ff-51143fcb650b	08/26/2014	Consumer Loan	...	Yes	No	1001740
2	9b2cd5d2-900e-4052-831f-6489f6d568af	08/22/2012	Consumer Loan	...	Yes	No	140039

[3 rows x 13 columns]

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- Load the data from csv file
- Display the data and showing the 3 data first
- The column or feature is 13

hanifETL > workETL.py > ...

```
10 conn = sqlite3.connect('user_data.db')
11 conn
12
13 # Write a query that selects everything from the users table.
14 query = 'SELECT * FROM users'
15 # Read the query into a DataFrame.
16 users = pd.read_sql(query, conn)
17 # Preview the data.
18 print(users.head());
19 # Check the shape of the data.
20 print(users.shape)
21
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
      user_id  age  job  marital  education  ...  pdays  previous  poutcome  term_deposit  date_joined
0  9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6  58  management  married  tertiary  ...  -1      0      None      no  1998-08-23
1  bb92765a-08de-4963-b432-496524b39157  44  technician  single  secondary  ...  -1      0      None      no  2008-07-15
2  573de577-49ef-42b9-83da-d3cfb817b5c1  33  entrepreneur  married  secondary  ...  -1      0      None      no  2002-06-04
3  d6b66b9d-7c8f-4257-a682-e136f640b7e3  47  blue-collar  married  None  ...  -1      0      None      no  1995-06-29
4  fade0b20-7594-4d9a-84cd-c02f79b1b526  33      None  single  None  ...  -1      0      None      no  1995-08-01

[5 rows x 16 columns]
(45216, 16)
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- Load the data from database user\_data.db
- select the users data
- Display data and showing the data by default 5 row only
- The column or feature is 16
- The total all data or rows is 45216

hanifETL > workETL.py > ...

```
22
23
24
25 query = 'SELECT * FROM device'
26 device = pd.read_sql(query, conn)
27 print(device.head())
28 print(device.shape)
29
30
31
32
33
...
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
      user_id  device
0  9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6  mobile
1  bb92765a-08de-4963-b432-496524b39157  desktop
2  573de577-49ef-42b9-83da-d3cfb817b5c1  mobile
3  d6b66b9d-7c8f-4257-a682-e136f640b7e3  tablet
4  fade0b20-7594-4d9a-84cd-c02f79b1b526  mobile
(45117, 2)
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- select the device data
- Display data and showing the data by default 5 row only
- The column or feature is 2
- The total all data or rows is 45117

```
workETL.py X
hanifETL > workETL.py > ...
31
32
33
34
35 # Read the user transactions in the last 30 days.
36 query = 'SELECT * FROM transactions'
37 transactions = pd.read_sql(query, conn)
38 print(transactions.head())
39 print(transactions.shape)
40
41
42
43
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
      user_id      transaction_id  amount_usd
0  9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6  transaction_5180      1332
1  9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6  transaction_5607       726
2  9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6  transaction_6765        85
3  573de577-49ef-42b9-83da-d3cfb817b5c1  transaction_6170         1
4  573de577-49ef-42b9-83da-d3cfb817b5c1  transaction_6090         1
(140034, 3)
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- select the transaction data
- Display data and showing the data by default 5 row only
- The column or feature is 3
- The total all data or rows is 140034

```
hanifETL > workETL.py > ...
45
46 # Aggregate data on the number of transactions and the total amount.
47 query = '''SELECT user_id,
48                COUNT(*) AS number_transactions,
49                SUM(amount_usd) AS total_amount_usd
50            FROM transactions
51            GROUP BY user_id'''
52
53 transactions_agg = pd.read_sql(query, conn)
54 print(transactions_agg.head())
55 print(transactions_agg.shape)
56
57
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
      user_id  number_transactions  total_amount_usd
0  0001570d-8aed-465e-b547-8981651084ed         3         792
1  000548ed-aa18-4eef-b8ed-68a9126e33ab         2        1044
2  00069959-4d55-460e-bb76-ae13ddb80a6         5          0
3  000bab00-aec4-4ee2-81a6-1f897c38726b        19          0
4  000cbac8-212f-46fb-b58f-861dada34284         2        399
(35211, 3)
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- Select data user with total transaction and total amount usd
- Display the data and showing the 3 data first
- Display data and showing the data by default 5 row only
- The total all data or rows is 35211

hanifETL > workETL.py > ...

```
61
62 # Do a left join, as all users in the users table are of interest.
63 query = '''SELECT left_table.*,
64                right_table.device
65                FROM users AS left_table
66                LEFT JOIN device AS right_table
67                ON left_table.user_id = right_table.user_id'''
68
69 users_w_device = pd.read_sql(query, conn)
70 print(users_w_device.head(n = 3))
71 print(users_w_device.shape)
72
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
```

	user_id	age	job	marital	education	...	previous	poutcome	term_deposit	date_joined	device
0	9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6	58	management	married	tertiary	...	0	None	no	1998-08-23	mobile
1	bb92765a-08de-4963-b432-496524b39157	44	technician	single	secondary	...	0	None	no	2008-07-15	desktop
2	573de577-49ef-42b9-83da-d3cfb817b5c1	33	entrepreneur	married	secondary	...	0	None	no	2002-06-04	mobile

[3 rows x 17 columns]

(45216, 17)

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> █

- Join 2 table user and device using sql query
- Display the data and showing the 3 data first
- The column or feature is 17
- The total all data or rows is 45216

```
76
77
78 # Do a right join so users won't be lost.
79 users_w_devices_and_transactions = \
80 transactions_agg.merge(users_w_device,
81                        on = 'user_id', how = 'right')
82
83 print(users_w_devices_and_transactions.head())
84 print(users_w_devices_and_transactions.shape)
85
86
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
```

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
0	9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6	3.0	2143.0	58	...	None	no	1998-08-23	mobile
1	bb92765a-08de-4963-b432-496524b39157	NaN	NaN	44	...	None	no	2008-07-15	desktop
2	573de577-49ef-42b9-83da-d3cfb817b5c1	2.0	2.0	33	...	None	no	2002-06-04	mobile
3	d6b66b9d-7c8f-4257-a682-e136f640b7e3	NaN	NaN	47	...	None	no	1995-06-29	tablet
4	fade0b20-7594-4d9a-84cd-c02f79b1b526	1.0	1.0	33	...	None	no	1995-08-01	mobile

[5 rows x 19 columns]

(45216, 19)

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> █

- Merge data user with device and transaction
- Display the data and showing the data by default
- The column or feature is 19
- The total all data or rows is 45216

## 2)Transform Data

```

88
89
90
91 #Identify data where age is greater than 150
92 print(users_w_devices_and_transactions[users_w_devices_and_transactions.age > 150])
93
94
95
96
97
98

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
7228	44fefdad-7045-4be5-890e-12e84ae6fdc9	NaN	NaN	178	...	None	no	1997-04-28	desktop
10318	9b2cd5d2-900e-4052-831f-6489f6d568af	2.0	3165.0	891	...	None	no	2003-10-03	mobile

[2 rows x 19 columns]

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> █

- Check the data user age is more than 150
- The column or feature is 19
- The total all data or rows is 2

```

98
99 #Drop incorrect data
100 users_cleaned = \
101 users_w_devices_and_transactions[users_w_devices_and_transactions.age < 150]
102
103 print(users_cleaned.shape)
104
105
106
107

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"

(45214, 19)

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> █

- Remove data which user age is more 150 or keep data user age below 150
- The column or feature is 19
- The original total all data is 45216 to 45214

```
hanifETL > workETL.py > ...
105
106
107
108
109 #Identify more potentially erroneous data
110 print(pd.crosstab(users_cleaned['age'], users_cleaned['device']))
111
112
113
114
115
116
117
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
device desktop mobile tablet
age
18      5      6      1
19     10     22      3
20     11     33      6
21     16     44     19
22     30     87     11
...
90      1      1      0
92      1      1      0
93      0      2      0
94      0      1      0
95      0      1      1

[77 rows x 3 columns]
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- Do crosstab data from row to column/feature such device [desktop, mobile, tablet]
- Display the total device used based on age data
- the age, desktop, mobile and tablet as feature
- The column or feature is 3
- The total all data or rows is 77

```
hanifETL > workETL.py > ...
118 crosstab_result = pd.crosstab(users_cleaned['age'], users_cleaned['device']);
119
120 device_counts = crosstab_result.sum(axis=1)
121
122 min_age = device_counts.idxmin()
123 max_age = device_counts.idxmax()
124
125 min_count = device_counts.min()
126 max_count = device_counts.max()
127
128 print(f"Age with the minimum total device usage: {min_age} ({min_count} devices)")
129 print(f"Age with the maximum total device usage: {max_age} ({max_count} devices)")
130
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
Age with the minimum total device usage: 94 (1 devices)
Age with the maximum total device usage: 32 (2082 devices)
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
```

- Find the maximum and minimum total device used by user based on age

```

hanifETL > workETL.py > ...
118 crosstab_result = pd.crosstab(users_cleaned['age'], users_cleaned['device']);
119
120 device_counts = crosstab_result.sum(axis=1)
121
122 min_age = device_counts.idxmin()
123 max_age = device_counts.idxmax()
124
125 min_count = device_counts.min()
126 max_count = device_counts.max()
127
128 #print(f"Age with the minimum total device usage: {min_age} ({min_count} devices)")
129 #print(f"Age with the maximum total device usage: {max_age} ({max_count} devices)")
130
131 # Device type data based on minimum and maximum age
132 devices_min_age = crosstab_result.loc[min_age]
133 devices_max_age = crosstab_result.loc[max_age]
134
135 # Types of devices with usage count for minimum and maximum age
136 print(f"Device types for age {min_age} (minimum usage):\n{devices_min_age}")
137 print(f"\nDevice types for age {max_age} (maximum usage):\n{devices_max_age}")
138

```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** PORTS

```

anif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
Device types for age 94 (minimum usage):
device
desktop    0
mobile     1
tablet     0
Name: 94, dtype: int64

Device types for age 32 (maximum usage):
device
desktop    557
mobile    1277
tablet     248
Name: 32, dtype: int64
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

```

- Find the maximum and minimum total for each device used by user based on age

```

workETL.py
hanifETL > workETL.py > ...
154 print(users_cleaned.info())
155 print(users_cleaned.default.value_counts())
156 print(users_cleaned.device.value_counts())
157

```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** PORTS

```

1 number_transactions 35215 non-null float64
2 total_amount_usd    35215 non-null float64
3 age                 45214 non-null int64
4 job                 44926 non-null object
5 marital             45214 non-null object
6 education           43357 non-null object
7 default             45214 non-null object
8 housing             45214 non-null object
9 loan                45214 non-null object
10 contact            32196 non-null object
11 duration           45214 non-null int64
12 campaign           45214 non-null int64
13 pdays             45214 non-null int64
14 previous           45214 non-null int64
15 poutcome           8255 non-null object
16 term_deposit       45214 non-null object
17 date_joined        45184 non-null object
18 device             45120 non-null object
dtypes: float64(2), int64(5), object(12)
memory usage: 6.9+ MB
None
default
no      44398
yes      816
Name: count, dtype: int64
device
mobile    27507
desktop   12114
tablet     5499
Name: count, dtype: int64
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

```

- Get the data info about structure user table like feature and type of data
- Total data of value for default feature
- Total data of value for device feature

```

159 #Convert the relevant variables to a Boolean type
160 users_cleaned_1 = users_cleaned.copy() # Work with a new object.
161
162 users_cleaned_1.default = \
163 users_cleaned_1.default.map(dict(yes = 1, no = 0)).astype(bool)
164
165 print(users_cleaned_1.default.value_counts())
166 print(users_cleaned_1.head())
167 print(users_cleaned_1["default"].head())

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & c:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
default
False    44398
True      816
Name: count, dtype: int64

```

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
0	9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6	3.0	2143.0	58	...	None	no	1998-08-23	mobile
1	bb92765a-08de-4963-b432-496524b39157	NaN	NaN	44	...	None	no	2008-07-15	desktop
2	573de577-49ef-42b9-83da-d3cfb817b5c1	2.0	2.0	33	...	None	no	2002-06-04	mobile
3	d6b66b9d-7c8f-4257-a682-e136f640b7e3	NaN	NaN	47	...	None	no	1995-06-29	tablet
4	fade0b20-7594-4d9a-84cd-c02f79b1b526	1.0	1.0	33	...	None	no	1995-08-01	mobile

```

[5 rows x 19 columns]
0    False
1    False
2    False
3    False
4    False
Name: default, dtype: bool
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

```

- Convert the default value to bool since only yes and no as value
- Display the data and showing the data by default
- Show feature default value



```
hanifETL > workETL.py > ...
171
172 # Do the same for the other Boolean variables.
173 bool_vars = ['housing', 'loan', 'term_deposit']
174 for var in bool_vars:
175     users_cleaned_1[var] = \
176     users_cleaned_1[var].map(dict(yes = 1, no = 0)).astype(bool)
177     print(f'Converted {var} to Boolean.')
178
179 print(users_cleaned_1.info())
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Converted housing to Boolean.  
Converted loan to Boolean.  
Converted term\_deposit to Boolean.  
<class 'pandas.core.frame.DataFrame'>  
Index: 45214 entries, 0 to 45215  
Data columns (total 19 columns):  
# Column Non-Null Count Dtype  
---  
0 user\_id 45214 non-null object  
1 number\_transactions 35215 non-null float64  
2 total\_amount\_usd 35215 non-null float64  
3 age 45214 non-null int64  
4 job 44926 non-null object  
5 marital 45214 non-null object  
6 education 43357 non-null object  
7 default 45214 non-null bool  
8 housing 45214 non-null bool  
9 loan 45214 non-null bool  
10 contact 32196 non-null object  
11 duration 45214 non-null int64  
12 campaign 45214 non-null int64  
13 pdays 45214 non-null int64  
14 previous 45214 non-null int64  
15 poutcome 8255 non-null object  
16 term\_deposit 45214 non-null bool  
17 date\_joined 45184 non-null object  
18 device 45120 non-null object  
dtypes: bool(4), float64(2), int64(5), object(8)

- Convert the housing, loan and term\_deposit value to bool format

```
hanifETL > workETL.py > ...
183 users_cleaned_2 = users_cleaned_1.copy() # Work with a new object.
184
185 users_cleaned_2['date_joined'] = \
186 pd.to_datetime(users_cleaned_2['date_joined'],
187               format = '%Y-%m-%d')
188
189 print(users_cleaned_1['date_joined'].head())
190 print(users_cleaned_2['date_joined'].head())
191 print(users_cleaned_1.info())
192 print(users_cleaned_2.info())
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Data columns (total 19 columns):  
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"  
0 1998-08-23  
1 2008-07-15  
2 2002-06-04  
3 1995-06-29  
4 1995-08-01  
Name: date\_joined, dtype: object  
0 1998-08-23  
1 2008-07-15  
2 2002-06-04  
3 1995-06-29  
4 1995-08-01  
Name: date\_joined, dtype: datetime64[ns]

- Convert the date\_joined value to date format as Y-M-D
- Show feature date\_joined value

```
Name: date_joined, dtype: datetime64[ns]
<class 'pandas.core.frame.DataFrame'>
Index: 45214 entries, 0 to 45215
Data columns (total 19 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   user_id                45214 non-null  object
1   number_transactions    35215 non-null  float64
2   total_amount_usd      35215 non-null  float64
3   age                    45214 non-null  int64
4   job                    44926 non-null  object
5   marital                45214 non-null  object
6   education              43357 non-null  object
7   default                45214 non-null  bool
8   housing                45214 non-null  bool
9   loan                   45214 non-null  bool
10  contact                32196 non-null  object
11  duration               45214 non-null  int64
12  campaign               45214 non-null  int64
13  pdays                 45214 non-null  int64
14  previous               45214 non-null  int64
15  poutcome               8255 non-null   object
16  term_deposit           45214 non-null  bool
17  date_joined            45184 non-null  object
18  device                 45120 non-null  object
dtypes: bool(4), float64(2), int64(5), object(8)
```

- Check data format for date\_joined before - object

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Index: 45214 entries, 0 to 45215

Data columns (total 19 columns):

#	Column	Non-Null	Count	Dtype
0	user_id	45214	non-null	object
1	number_transactions	35215	non-null	float64
2	total_amount_usd	35215	non-null	float64
3	age	45214	non-null	int64
4	job	44926	non-null	object
5	marital	45214	non-null	object
6	education	43357	non-null	object
7	default	45214	non-null	bool
8	housing	45214	non-null	bool
9	loan	45214	non-null	bool
10	contact	32196	non-null	object
11	duration	45214	non-null	int64
12	campaign	45214	non-null	int64
13	pdays	45214	non-null	int64
14	previous	45214	non-null	int64
15	poutcome	8255	non-null	object
16	term_deposit	45214	non-null	bool
17	date_joined	45184	non-null	datetime64[ns]
18	device	45120	non-null	object

dtypes: bool(4), datetime64[ns](1), float64(2), int64(5), object(7)  
memory usage: 5.7+ MB  
None

- Check data format for date\_joined after – datetime64[ns]

```
hanifETL > workETL.py > ...
195 #Identify all duplicated data
196 duplicated_data = \
197 users_cleaned_2[users_cleaned_2.duplicated(keep = False)]
198
199 duplicated_data = \
200 duplicated_data.sort_values(by = list(duplicated_data.columns))
201
202 print('Number of rows with duplicated data:',
203       duplicated_data.shape[0])
204
205 print(duplicated_data)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
```

```
Number of rows with duplicated data: 10
      user_id  number_transactions  total_amount_usd  age  ...  poutcome  term_deposit  date_joined  device
22005  1e826721-b38c-41c2-88f4-4c28b335b1e6          4.0    159.0  31  ...    None          False  1990-05-10  mobile
45212  1e826721-b38c-41c2-88f4-4c28b335b1e6          4.0    159.0  31  ...    None          False  1990-05-10  mobile
35413  a2fb8264-d55a-437b-a8e7-9ec4116b76f4          2.0    676.0  34  ...  success          False  1989-09-02  mobile
45213  a2fb8264-d55a-437b-a8e7-9ec4116b76f4          2.0    676.0  34  ...  success          False  1989-09-02  mobile
36292  ae3b92a2-cad8-434f-8037-9815e2228839          2.0    426.0  43  ...  failure          False  1992-10-29  desktop
45214  ae3b92a2-cad8-434f-8037-9815e2228839          2.0    426.0  43  ...  failure          False  1992-10-29  desktop
15456  cba59442-af3c-41d7-a39c-0f9bffa0660          2.0   1218.0  57  ...    None          False  1993-02-27  desktop
45211  cba59442-af3c-41d7-a39c-0f9bffa0660          2.0   1218.0  57  ...    None          False  1993-02-27  desktop
35620  f49ac08f-b872-4d57-ac82-9b8a9144020d          4.0    117.0  38  ...  failure          False  2010-06-11  mobile
45215  f49ac08f-b872-4d57-ac82-9b8a9144020d          4.0    117.0  38  ...  failure          False  2010-06-11  mobile
```

[10 rows x 19 columns]

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> 
```

- Check duplicate data
- The total all data or rows is 10

hanifETL > workETL.py > ...

```
215 #Remove the duplicated data
216 users_cleaned_final = \
217     users_cleaned_2[~users_cleaned_2.duplicated()]
218
219 print(users_cleaned_2[users_cleaned_2['user_id'] == "cba59442-af3c-41d7-a39c-0f9bffb0660"])
220 print(users_cleaned_final[users_cleaned_final['user_id'] == "cba59442-af3c-41d7-a39c-0f9bffb0660"])
221
222 print(users_cleaned_2.shape)
223 print(users_cleaned_final.shape)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

anif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
22005	1e826721-b38c-41c2-88f4-4c28b335b1e6	4.0	159.0	31	...	None	False	1990-05-10	mobile
45212	1e826721-b38c-41c2-88f4-4c28b335b1e6	4.0	159.0	31	...	None	False	1990-05-10	mobile
35413	a2fb8264-d55a-437b-a8e7-9ec4116b76f4	2.0	676.0	34	...	success	False	1989-09-02	mobile
45213	a2fb8264-d55a-437b-a8e7-9ec4116b76f4	2.0	676.0	34	...	success	False	1989-09-02	mobile
36292	ae3b92a2-cad8-434f-8037-9815e2228839	2.0	426.0	43	...	failure	False	1992-10-29	desktop
45214	ae3b92a2-cad8-434f-8037-9815e2228839	2.0	426.0	43	...	failure	False	1992-10-29	desktop
15456	cba59442-af3c-41d7-a39c-0f9bffb0660	2.0	1218.0	57	...	None	False	1993-02-27	desktop
45211	cba59442-af3c-41d7-a39c-0f9bffb0660	2.0	1218.0	57	...	None	False	1993-02-27	desktop
35620	f49ac08f-b872-4d57-ac82-9b8a9144020d	4.0	117.0	38	...	failure	False	2010-06-11	mobile
45215	f49ac08f-b872-4d57-ac82-9b8a9144020d	4.0	117.0	38	...	failure	False	2010-06-11	mobile

[10 rows x 19 columns]

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
15456	cba59442-af3c-41d7-a39c-0f9bffb0660	2.0	1218.0	57	...	None	False	1993-02-27	desktop
45211	cba59442-af3c-41d7-a39c-0f9bffb0660	2.0	1218.0	57	...	None	False	1993-02-27	desktop

[2 rows x 19 columns]

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
15456	cba59442-af3c-41d7-a39c-0f9bffb0660	2.0	1218.0	57	...	None	False	1993-02-27	desktop

[1 rows x 19 columns]

(45214, 19)

(45209, 19)

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> a2fb8264-d55a-437b-a8e7-9ec4116b76f4

- Remove duplicate data
- Display the data and showing the data based on user id
- Check total data before and after remove duplicate
- Before – 45214, after - 45209

### 3)Load Data

```
226
227 #Load data into an SQL database
228 conn = sqlite3.connect('users_data_cleaned.db')
229
230 users_cleaned_final.to_sql('users_cleaned_final',
231                             conn,
232                             if_exists = 'replace',
233                             index = False)
234
235 query = 'SELECT * FROM users_cleaned_final'
236
237 datafinal = pd.read_sql(query, conn).head()
238 print(datafinal.head())
239 conn.close()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:\Users\hanif\AppData\Local\Programs\Python\Python312\python.exe "c:/Users/hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"

	user_id	number_transactions	total_amount_usd	age	...	poutcome	term_deposit	date_joined	device
0	9231c446-cb16-4b2b-a7f7-ddfc8b25aaf6	3.0	2143.0	58	...	None	0	1998-08-23 00:00:00	mobile
1	bb92765a-08de-4963-b432-496524b39157	NaN	NaN	44	...	None	0	2008-07-15 00:00:00	desktop
2	573de577-49ef-42b9-83da-d3cfb817b5c1	2.0	2.0	33	...	None	0	2002-06-04 00:00:00	mobile
3	d6b66b9d-7c8f-4257-a682-e136f6a0b7e3	NaN	NaN	47	...	None	0	1995-06-29 00:00:00	tablet
4	fade0b20-7594-4d9a-84cd-c02f79b1b526	1.0	1.0	33	...	None	0	1995-08-01 00:00:00	mobile

[5 rows x 19 columns]

PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

- Create db file for user data that has been cleaned

- Display the data and showing the data by default
- The column or feature is 19

```
hanifETL > workETL.py > ...
244 #Write the DataFrame as a pickle file
245 users_cleaned_final.to_pickle('users_data_cleaned.pickle')
246
247 #Confirm that the data was written to the pickle file
248 pd.read_pickle('users_data_cleaned.pickle').head()
249 pd.read_pickle('users_data_cleaned.pickle').info()
250
```

	PROBLEMS	OUTPUT	DEBUG CONSOLE	TERMINAL	PORTS
0	user_id	45209	non-null	object	
1	number_transactions	35210	non-null	float64	
2	total_amount_usd	35210	non-null	float64	
3	age	45209	non-null	int64	
4	job	44921	non-null	object	
5	marital	45209	non-null	object	
6	education	43352	non-null	object	
7	default	45209	non-null	bool	
8	housing	45209	non-null	bool	
9	loan	45209	non-null	bool	
10	contact	32191	non-null	object	
11	duration	45209	non-null	int64	
12	campaign	45209	non-null	int64	
13	pdays	45209	non-null	int64	
14	previous	45209	non-null	int64	
15	poutcome	8252	non-null	object	
16	term_deposit	45209	non-null	bool	
17	date_joined	45179	non-null	datetime64[ns]	
18	device	45115	non-null	object	

dtypes: bool(4), datetime64[ns](1), float64(2), int64(5), object(7)  
memory usage: 5.7+ MB  
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL>

- Create a pickle file
- Saves complex objects like **lists, dictionaries, tuples, or classes**
- Simplifies **data transfer** between systems
- Used in **Machine Learning** to save trained models
- Store the data that has been cleaned

```
hanifETL > workETL.py > ...

255 #Write the data to a CSV file
256 users_cleaned_final.to_csv('users_data_cleaned.csv',
257                             index = False)
258
259 #Confirm that the data was written to a CSV file.
260 pd.read_csv('users_data_cleaned.csv').head()
261 pd.read_csv('users_data_cleaned.csv').info()
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> & C:/Users/hanif/AppData/Local/Programs/Python/Python312/python.exe "c:/Users/
hanif/OneDrive/Desktop/DIGITAL LEADERSHIP/hanifETL/workETL.py"
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45209 entries, 0 to 45208
Data columns (total 19 columns):
#   Column                Non-Null Count  Dtype
---  -
0   user_id                45209 non-null  object
1   number_transactions    35210 non-null  float64
2   total_amount_usd      35210 non-null  float64
3   age                    45209 non-null  int64
4   job                    44921 non-null  object
5   marital                45209 non-null  object
6   education              43352 non-null  object
7   default                45209 non-null  bool
8   housing                45209 non-null  bool
9   loan                   45209 non-null  bool
10  contact                32191 non-null  object
11  duration               45209 non-null  int64
12  campaign               45209 non-null  int64
13  pdays                  45209 non-null  int64
14  previous               45209 non-null  int64
15  poutcome               8252 non-null   object
16  term_deposit           45209 non-null  bool
17  date_joined            45179 non-null  object
18  device                 45115 non-null  object
dtypes: bool(4), float64(2), int64(5), object(8)
memory usage: 5.3+ MB
PS C:\Users\hanif\OneDrive\Desktop\DIGITAL LEADERSHIP\hanifETL> 
```

- Create the csv file
- Store the data that has been cleaned