BWT Task-06 Exercise

Submitted By: ABRAR SAEED

ETL Task

There are some problems with the data and you need to remove them during the transformation phase to make data useful for everyone:

- An order id should always exist as an integer

- A product id cannot be 0

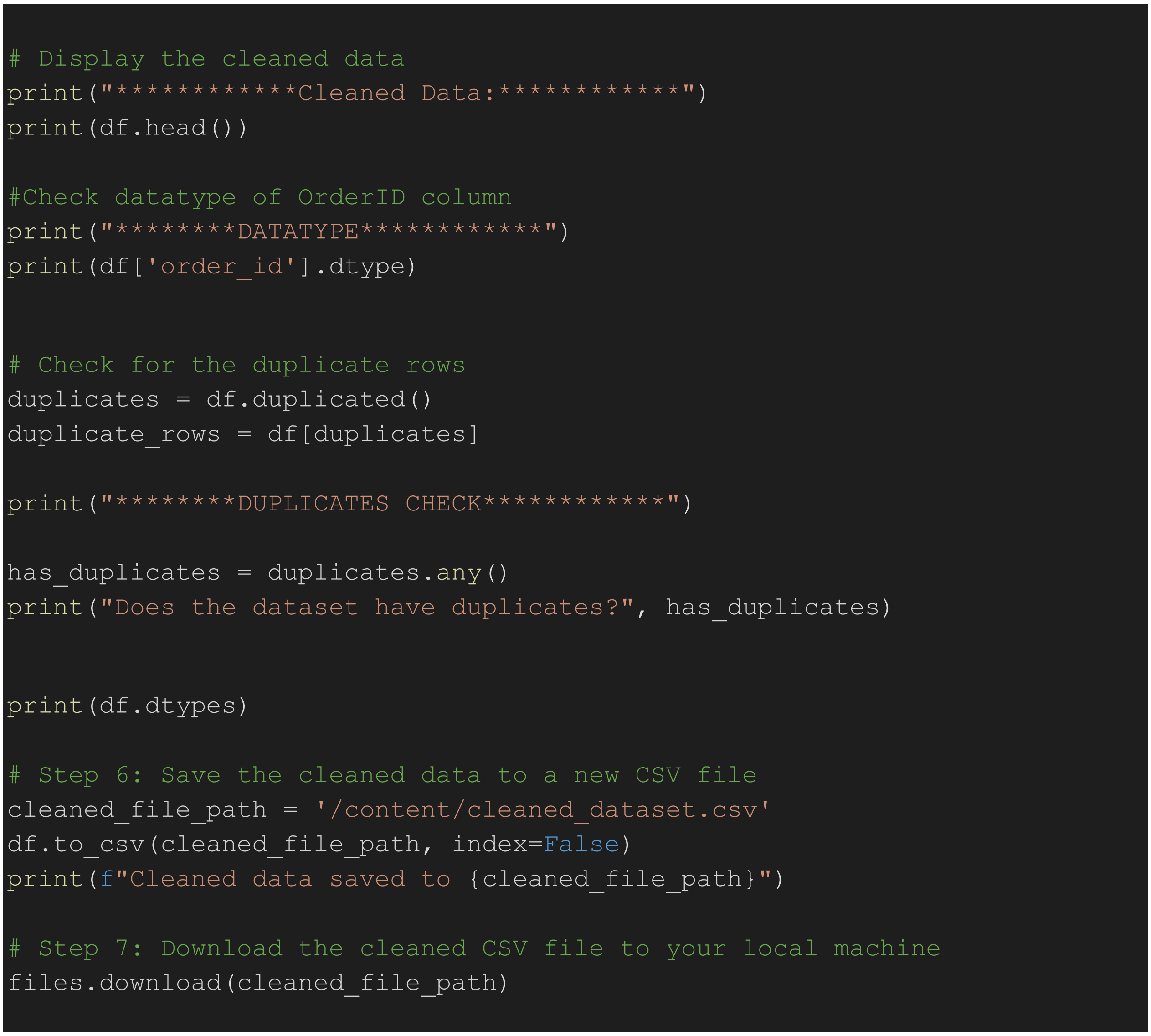
- We never had a product priced more than 1500 Rs. so any item with amount greater than 1500 Rs is an anomaly and it should be treated as 1500 Rs

- A status of an item can never be null or None, if it is then its an anomaly and item rows to be considered as fake orders and should not be kept in final data

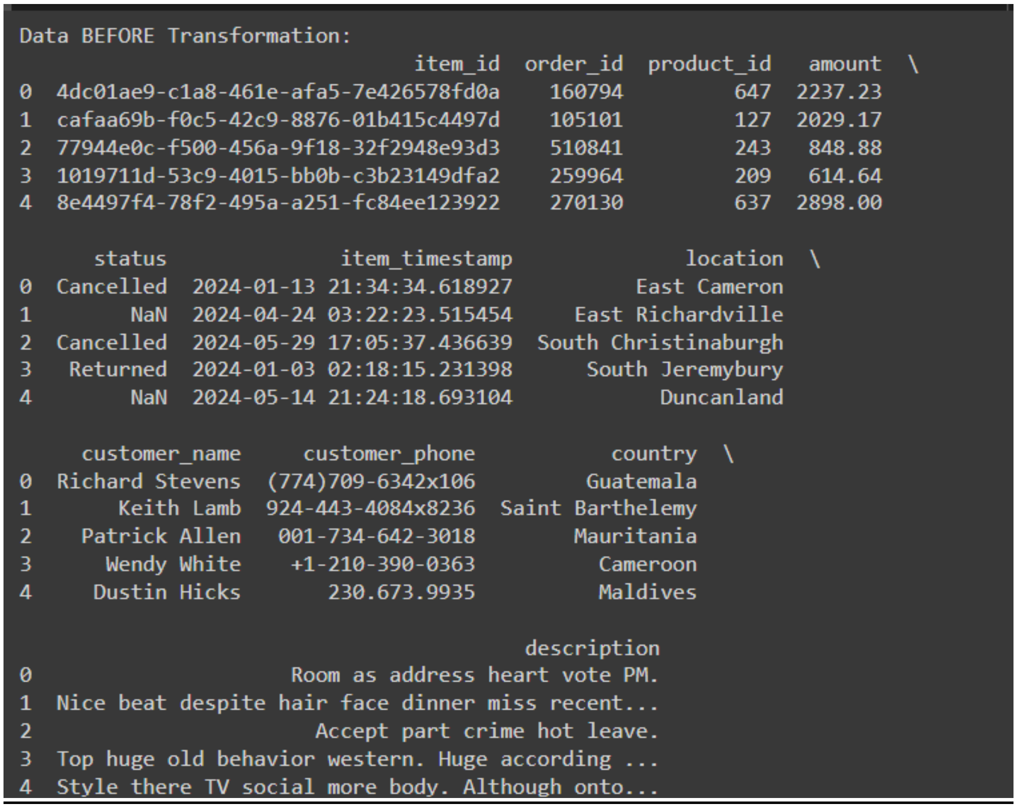
- There must be duplication in final data

Code:

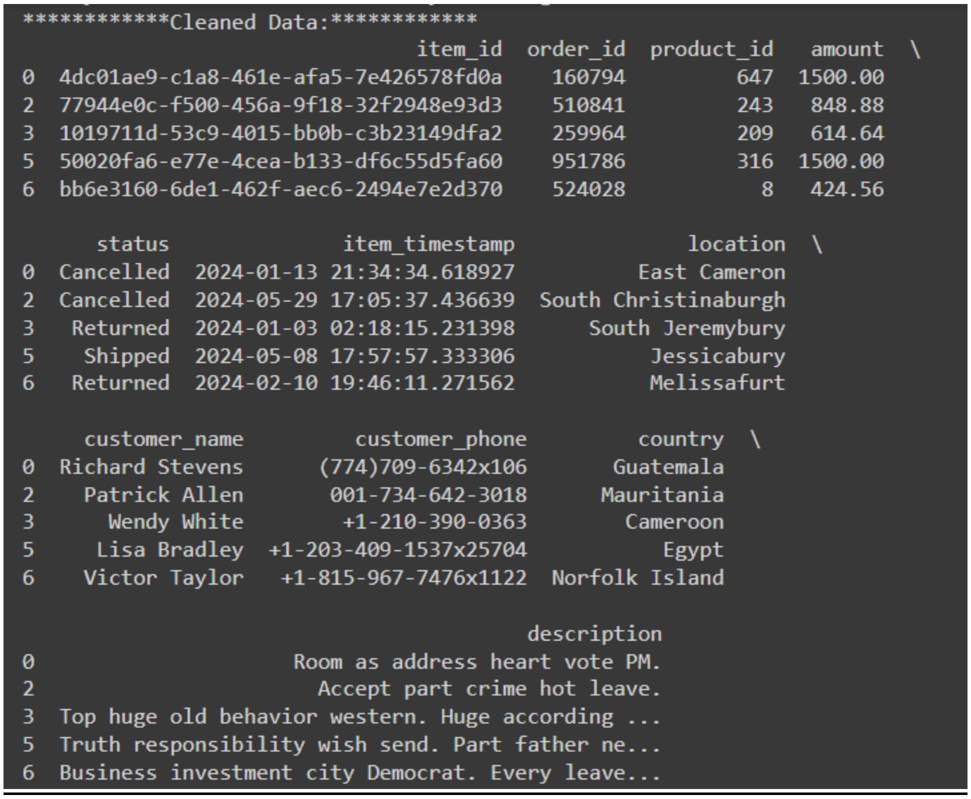
```
import pandas as pd
from google.colab import files

# Load the data
file_path = '/content/drive/MyDrive/dataset.csv'
df = pd.read_csv(file_path)

# Display the first few rows of the dataframe
print("Data BEFORE Transformation:")
print(df.head())

# 1. An order id should always exist as an integer
df = df.dropna(subset=['order_id'])
df['order_id'] = df['order_id'].astype(int)

# 2. A product id cannot be 0
df = df[df['product_id'] != 0]

# 3. We never had a product priced more than 1500 Rs. so any item with amount greater than 1500 Rs
#is an anomaly and it should be treated as 1500 Rs
df['amount'] = df['amount'].apply(lambda x: min(x, 1500))

# 4. Remove rows where `status` is null or None
df = df.dropna(subset=['status'])
```



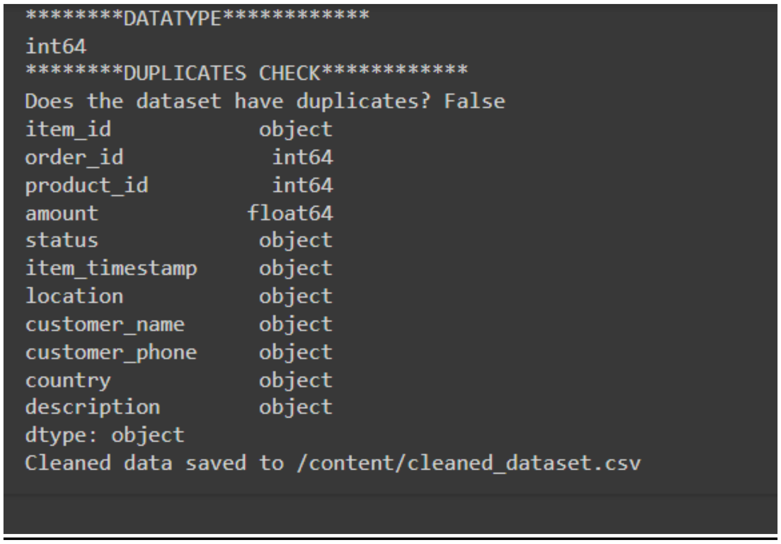
```
# Display the cleaned data
print("************Cleaned Data:************")
print(df.head())

#Check datatype of OrderID column
print("********DATATYPE************")
print(df['order_id'].dtype)


# Check for the duplicate rows
duplicates = df.duplicated()
duplicate_rows = df[duplicates]

print("********DUPLICATES CHECK************")

has_duplicates = duplicates.any()
print("Does the dataset have duplicates?", has_duplicates)


print(df.dtypes)

# Step 6: Save the cleaned data to a new CSV file
cleaned_file_path = '/content/cleaned_dataset.csv'
df.to_csv(cleaned_file_path, index=False)
print(f"Cleaned data saved to {cleaned_file_path}")

# Step 7: Download the cleaned CSV file to your local machine
files.download(cleaned_file_path)
```



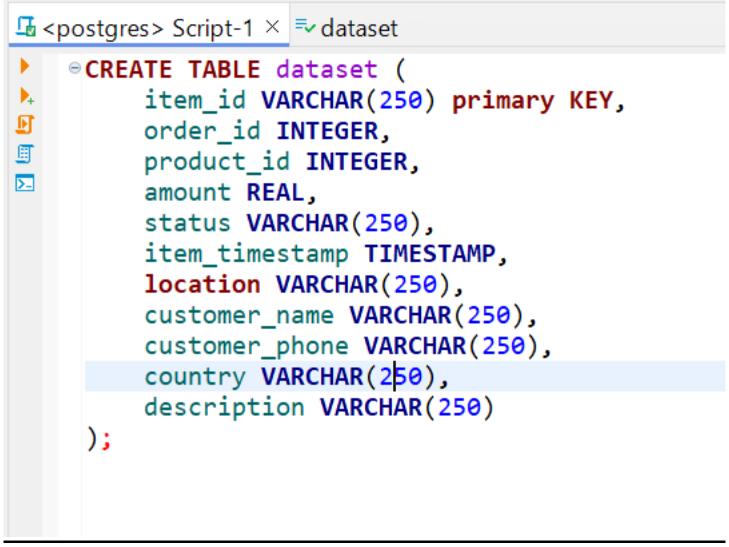
```
Data BEFORE Transformation:
                                item_id  order_id  product_id   amount  \
0  4dc01ae9-c1a8-461e-afa5-7e426578fd0a    160794         647  2237.23
1  cafaa69b-f0c5-42c9-8876-01b415c4497d    105101         127  2029.17
2  77944e0c-f500-456a-9f18-32f2948e93d3    510841         243   848.88
3  1019711d-53c9-4015-bb0b-c3b23149dfa2    259964         209   614.64
4  8e4497f4-78f2-495a-a251-fc84ee123922    270130         637  2898.00

      status              item_timestamp              location  \
0  Cancelled  2024-01-13 21:34:34.618927          East Cameron
1        NaN  2024-04-24 03:22:23.515454     East Richardville
2  Cancelled  2024-05-29 17:05:37.436639  South Christinaburgh
3   Returned  2024-01-03 02:18:15.231398       South Jeremybury
4        NaN  2024-05-14 21:24:18.693104            Duncanland

     customer_name     customer_phone            country  \
0  Richard Stevens  (774)709-6342x106          Guatemala
1       Keith Lamb  924-443-4084x8236  Saint Barthelemy
2    Patrick Allen    001-734-642-3018         Mauritania
3      Wendy White    +1-210-390-0363           Cameroon
4     Dustin Hicks       230.673.9935           Maldives

                                         description
0                       Room as address heart vote PM.
1  Nice beat despite hair face dinner miss recent...
2                         Accept part crime hot leave.
3  Top huge old behavior western. Huge according ...
4  Style there TV social more body. Although onto...
```



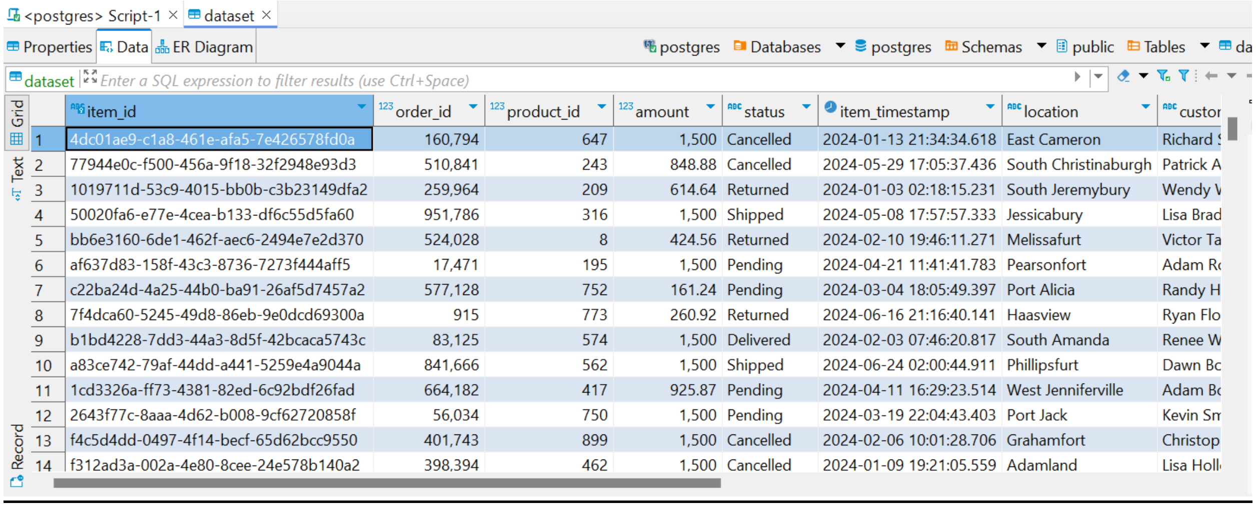
```
************Cleaned Data:************
                                item_id  order_id  product_id   amount  \
0  4dc01ae9-c1a8-461e-afa5-7e426578fd0a    160794         647  1500.00
2  77944e0c-f500-456a-9f18-32f2948e93d3    510841         243   848.88
3  1019711d-53c9-4015-bb0b-c3b23149dfa2    259964         209   614.64
5  50020fa6-e77e-4cea-b133-df6c55d5fa60    951786         316  1500.00
6  bb6e3160-6de1-462f-aec6-2494e7e2d370    524028           8   424.56

      status              item_timestamp              location  \
0  Cancelled  2024-01-13 21:34:34.618927          East Cameron
2  Cancelled  2024-05-29 17:05:37.436639  South Christinaburgh
3   Returned  2024-01-03 02:18:15.231398       South Jeremybury
5    Shipped  2024-05-08 17:57:57.333306            Jessicabury
6   Returned  2024-02-10 19:46:11.271562            Melissafurt

     customer_name         customer_phone          country  \
0  Richard Stevens      (774)709-6342x106        Guatemala
2    Patrick Allen       001-734-642-3018       Mauritania
3      Wendy White        +1-210-390-0363         Cameroon
5     Lisa Bradley  +1-203-409-1537x25704            Egypt
6    Victor Taylor   +1-815-967-7476x1122   Norfolk Island

                                         description
0                       Room as address heart vote PM.
2                         Accept part crime hot leave.
3  Top huge old behavior western. Huge according ...
5  Truth responsibility wish send. Part father ne...
6  Business investment city Democrat. Every leave...
```



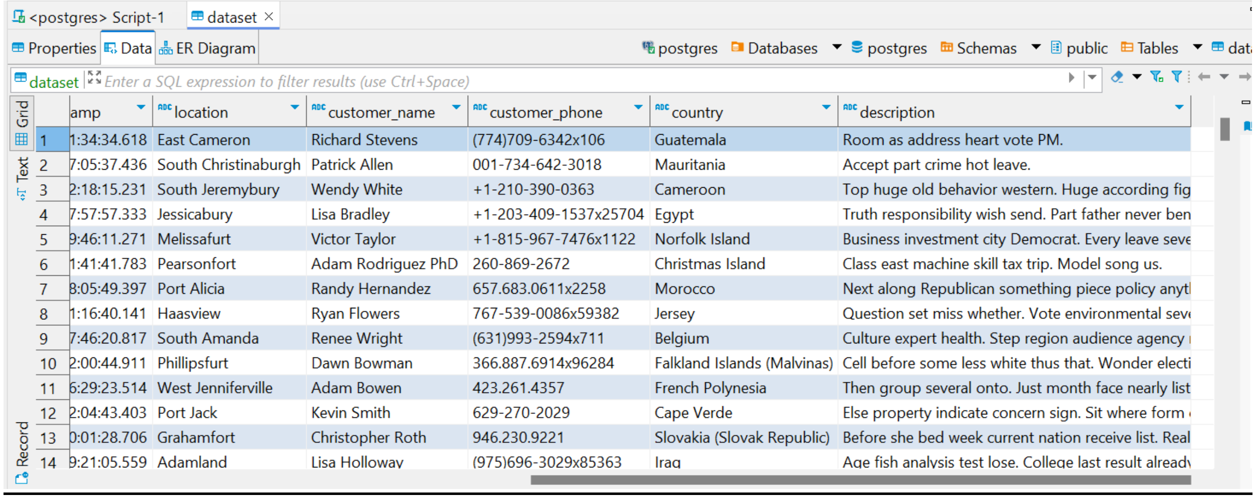
```
********DATATYPE************
int64
********DUPLICATES CHECK************
Does the dataset have duplicates? False
item_id             object
order_id             int64
product_id           int64
amount             float64
status              object
item_timestamp      object
location            object
customer_name       object
customer_phone      object
country             object
description         object
dtype: object
Cleaned data saved to /content/cleaned_dataset.csv
```



```sql
CREATE TABLE dataset (
    item_id VARCHAR(250) primary KEY,
    order_id INTEGER,
    product_id INTEGER,
    amount REAL,
    status VARCHAR(250),
    item_timestamp TIMESTAMP,
    location VARCHAR(250),
    customer_name VARCHAR(250),
    customer_phone VARCHAR(250),
    country VARCHAR(250),
    description VARCHAR(250)
);
```



	item_id	order_id	product_id	amount	status	item_timestamp	location	custor
1	4dc01ae9-c1a8-461e-afa5-7e426578fd0a	160,794	647	1,500	Cancelled	2024-01-13 21:34:34.618	East Cameron	Richard S
2	77944e0c-f500-456a-9f18-32f2948e93d3	510,841	243	848.88	Cancelled	2024-05-29 17:05:37.436	South Christinaburgh	Patrick A
3	1019711d-53c9-4015-bb0b-c3b23149dfa2	259,964	209	614.64	Returned	2024-01-03 02:18:15.231	South Jeremybury	Wendy V
4	50020fa6-e77e-4cea-b133-df6c55d5fa60	951,786	316	1,500	Shipped	2024-05-08 17:57:57.333	Jessicabury	Lisa Brad
5	bb6e3160-6de1-462f-aec6-2494e7e2d370	524,028	8	424.56	Returned	2024-02-10 19:46:11.271	Melissafurt	Victor Ta
6	af637d83-158f-43c3-8736-7273f444aff5	17,471	195	1,500	Pending	2024-04-21 11:41:41.783	Pearsonfort	Adam Ro
7	c22ba24d-4a25-44b0-ba91-26af5d7457a2	577,128	752	161.24	Pending	2024-03-04 18:05:49.397	Port Alicia	Randy H
8	7f4dca60-5245-49d8-86eb-9e0dcd69300a	915	773	260.92	Returned	2024-06-16 21:16:40.141	Haasview	Ryan Flo
9	b1bd4228-7dd3-44a3-8d5f-42bcaca5743c	83,125	574	1,500	Delivered	2024-02-03 07:46:20.817	South Amanda	Renee W
10	a83ce742-79af-44dd-a441-5259e4a9044a	841,666	562	1,500	Shipped	2024-06-24 02:00:44.911	Phillipsfurt	Dawn Bc
11	1cd3326a-ff73-4381-82ed-6c92bdf26fad	664,182	417	925.87	Pending	2024-04-11 16:29:23.514	West Jenniferville	Adam Bc
12	2643f77c-8aaa-4d62-b008-9cf62720858f	56,034	750	1,500	Pending	2024-03-19 22:04:43.403	Port Jack	Kevin Sm
13	f4c5d4dd-0497-4f14-becf-65d62bcc9550	401,743	899	1,500	Cancelled	2024-02-06 10:01:28.706	Grahamfort	Christop
14	f312ad3a-002a-4e80-8cee-24e578b140a2	398,394	462	1,500	Cancelled	2024-01-09 19:21:05.559	Adamland	Lisa Holl



	amp	location	customer_name	customer_phone	country	description
1	1:34:34.618	East Cameron	Richard Stevens	(774)709-6342x106	Guatemala	Room as address heart vote PM.
2	7:05:37.436	South Christinaburgh	Patrick Allen	001-734-642-3018	Mauritania	Accept part crime hot leave.
3	2:18:15.231	South Jeremybury	Wendy White	+1-210-390-0363	Cameroon	Top huge old behavior western. Huge according fig
4	7:57:57.333	Jessicabury	Lisa Bradley	+1-203-409-1537x25704	Egypt	Truth responsibility wish send. Part father never ben
5	9:46:11.271	Melissafurt	Victor Taylor	+1-815-967-7476x1122	Norfolk Island	Business investment city Democrat. Every leave seve
6	1:41:41.783	Pearsonfort	Adam Rodriguez PhD	260-869-2672	Christmas Island	Class east machine skill tax trip. Model song us.
7	8:05:49.397	Port Alicia	Randy Hernandez	657.683.0611x2258	Morocco	Next along Republican something piece policy anyt
8	1:16:40.141	Haasview	Ryan Flowers	767-539-0086x59382	Jersey	Question set miss whether. Vote environmental seve
9	7:46:20.817	South Amanda	Renee Wright	(631)993-2594x711	Belgium	Culture expert health. Step region audience agency
10	2:00:44.911	Phillipsfurt	Dawn Bowman	366.887.6914x96284	Falkland Islands (Malvinas)	Cell before some less white thus that. Wonder electi
11	6:29:23.514	West Jenniferville	Adam Bowen	423.261.4357	French Polynesia	Then group several onto. Just month face nearly list
12	2:04:43.403	Port Jack	Kevin Smith	629-270-2029	Cape Verde	Else property indicate concern sign. Sit where form
13	0:01:28.706	Grahamfort	Christopher Roth	946.230.9221	Slovakia (Slovak Republic)	Before she bed week current nation receive list. Real
14	9:21:05.559	Adamland	Lisa Holloway	(975)696-3029x85363	Iraq	Age fish analysis test lose. College last result already

Explanation:

A CSV file was given for which the ETL tasks had to be performed which in this instance were to read data from dataset.csv file using pandas or PySpark. But in the above mentioned code, the data was extracted using pandas and then the necessary transformations were carried out on the csv based on the task requirements and the data was loaded in the database to view the transformed data and to use it further.