ELEVATE LABS

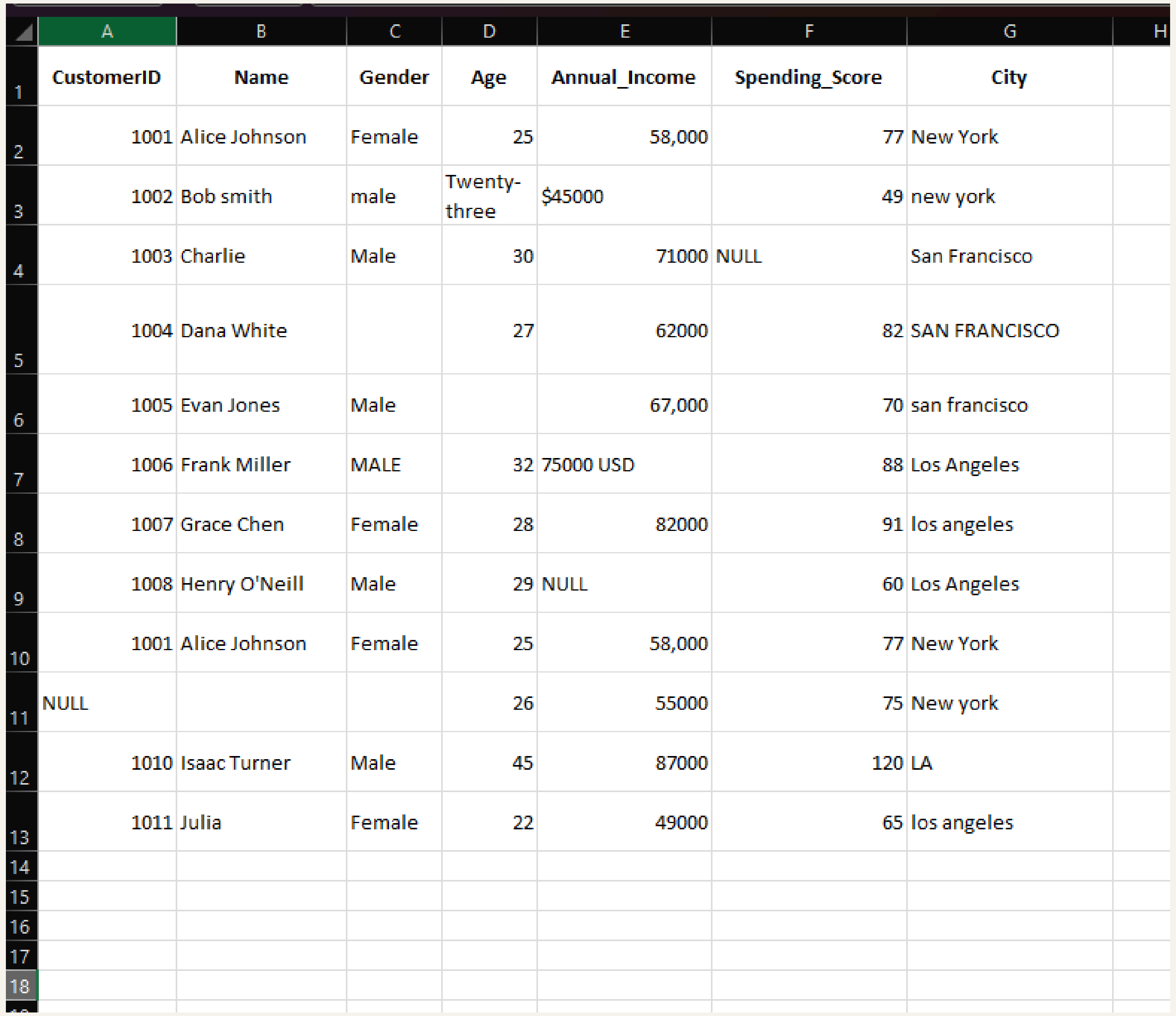
Project 1

TAKING FIRST STEPS TO CLEANING DATASET

BY - ARJUN JAISWAL

Raw Mall Customer Segmentation Dataset (with issues)

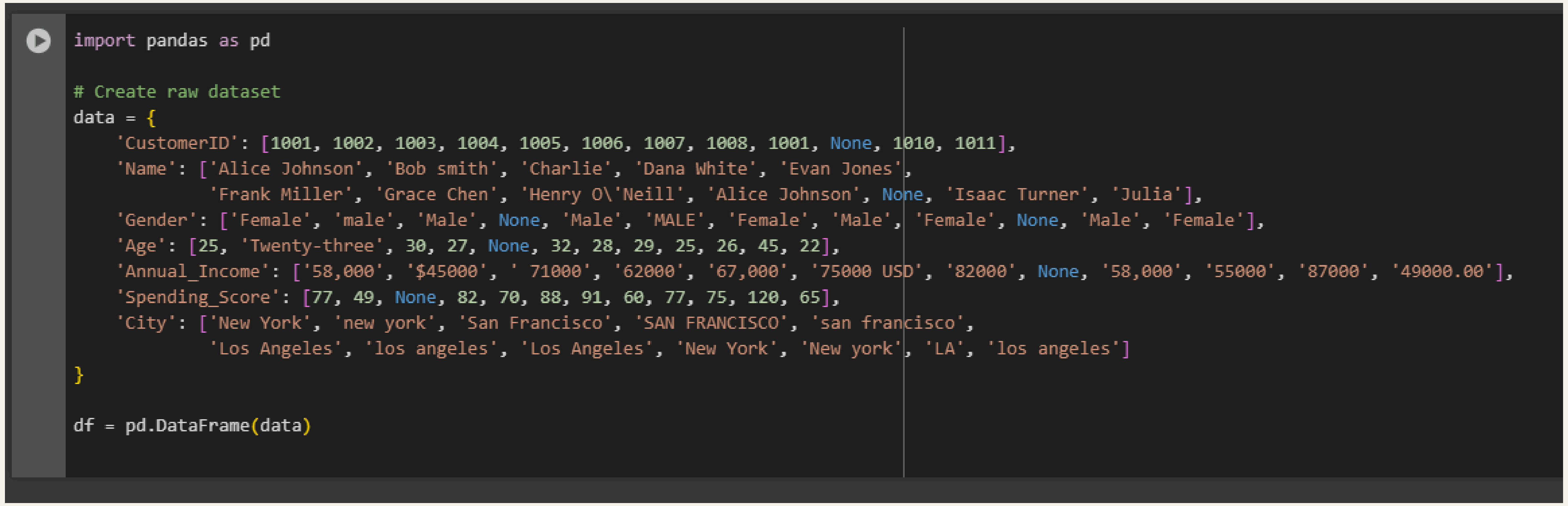
This is the Mall customer dataset that I will be working on to clean it and make it presentable as per the assignment.



CustomerID	Name	Gender	Age	Annual_Income	Spending_Score	City
1001	Alice Johnson	Female	25	58,000	77	New York
1002	Bob smith	male	Twenty-three	$45000	49	new york
1003	Charlie	Male	30	71000	NULL	San Francisco
1004	Dana White		27	62000	82	SAN FRANCISCO
1005	Evan Jones	Male		67,000	70	san francisco
1006	Frank Miller	MALE	32	75000 USD	88	Los Angeles
1007	Grace Chen	Female	28	82000	91	los angeles
1008	Henry O'Neill	Male	29	NULL	60	Los Angeles
1001	Alice Johnson	Female	25	58,000	77	New York
NULL			26	55000	75	New york
1010	Isaac Turner	Male	45	87000	120	LA
1011	Julia	Female	22	49000	65	los angeles

Identifying and handling missing value(using python)

Importing all the values as given in raw dataset and importing panda to carry on with further functions.

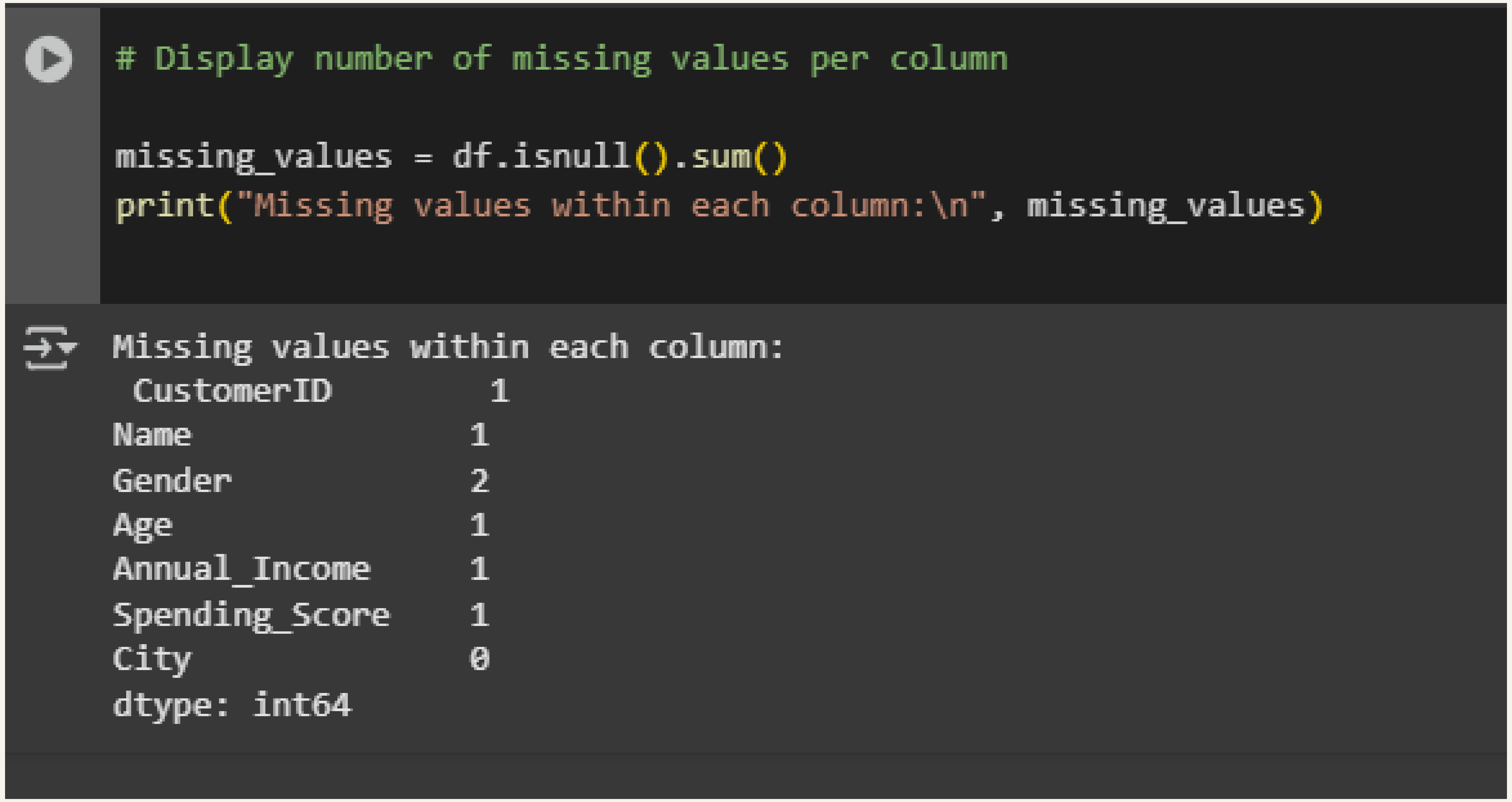


```python
import pandas as pd

# Create raw dataset
data = {
    'CustomerID': [1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1001, None, 1010, 1011],
    'Name': ['Alice Johnson', 'Bob smith', 'Charlie', 'Dana White', 'Evan Jones',
             'Frank Miller', 'Grace Chen', 'Henry O\'Neill', 'Alice Johnson', None, 'Isaac Turner', 'Julia'],
    'Gender': ['Female', 'male', 'Male', None, 'Male', 'MALE', 'Female', 'Male', 'Female', None, 'Male', 'Female'],
    'Age': [25, 'Twenty-three', 30, 27, None, 32, 28, 29, 25, 26, 45, 22],
    'Annual_Income': ['58,000', '$45000', ' 71000', '62000', '67,000', '75000 USD', '82000', None, '58,000', '55000', '87000', '49000.00'],
    'Spending_Score': [77, 49, None, 82, 70, 88, 91, 60, 77, 75, 120, 65],
    'City': ['New York', 'new york', 'San Francisco', 'SAN FRANCISCO', 'san francisco',
             'Los Angeles', 'los angeles', 'Los Angeles', 'New York', 'New york', 'LA', 'los angeles']
}

df = pd.DataFrame(data)
```

All missing values identified using isnull() function

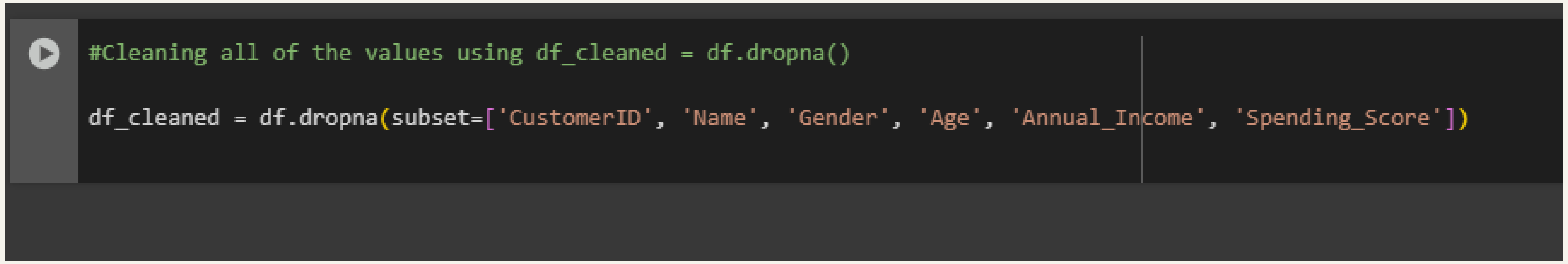


```
# Display number of missing values per column

missing_values = df.isnull().sum()
print("Missing values within each column:\n", missing_values)
```

```
Missing values within each column:
 CustomerID        1
Name              1
Gender            2
Age               1
Annual_Income     1
Spending_Score    1
City              0
dtype: int64
```

All of the values are successfully cleaned using df_cleaned = df.dropna(subset)

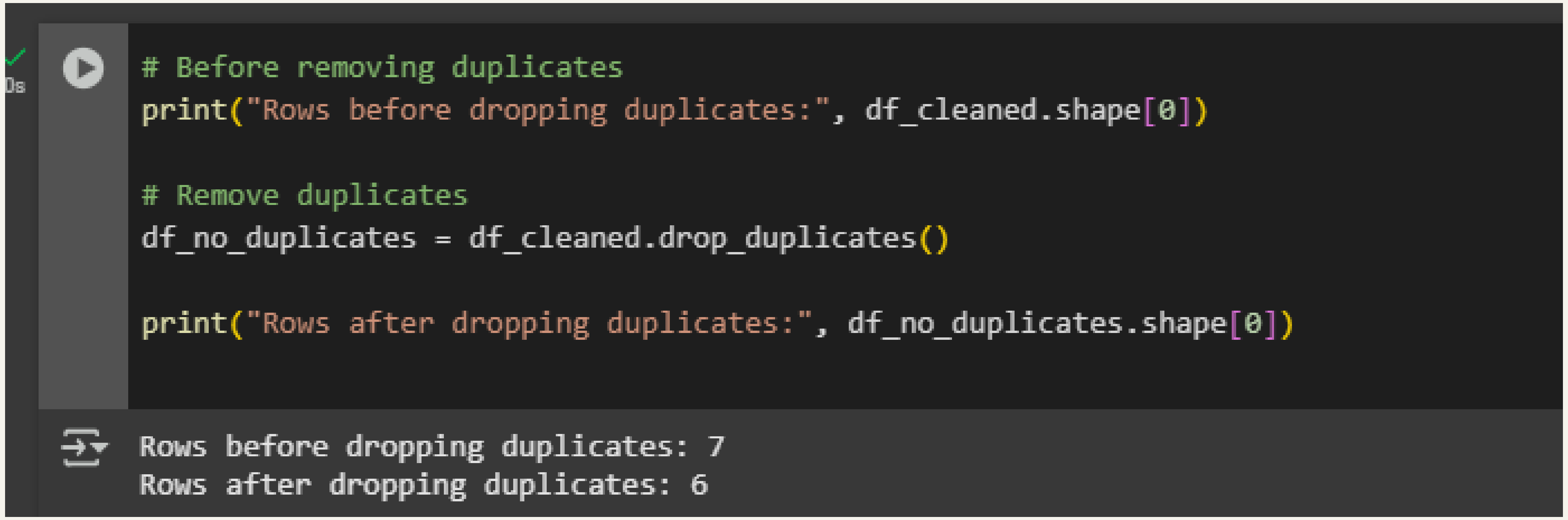


```
#Cleaning all of the values using df_cleaned = df.dropna()

df_cleaned = df.dropna(subset=['CustomerID', 'Name', 'Gender', 'Age', 'Annual_Income', 'Spending_Score'])
```

Made sure only those columns
are used which have null values

Cleaned all of the duplicates



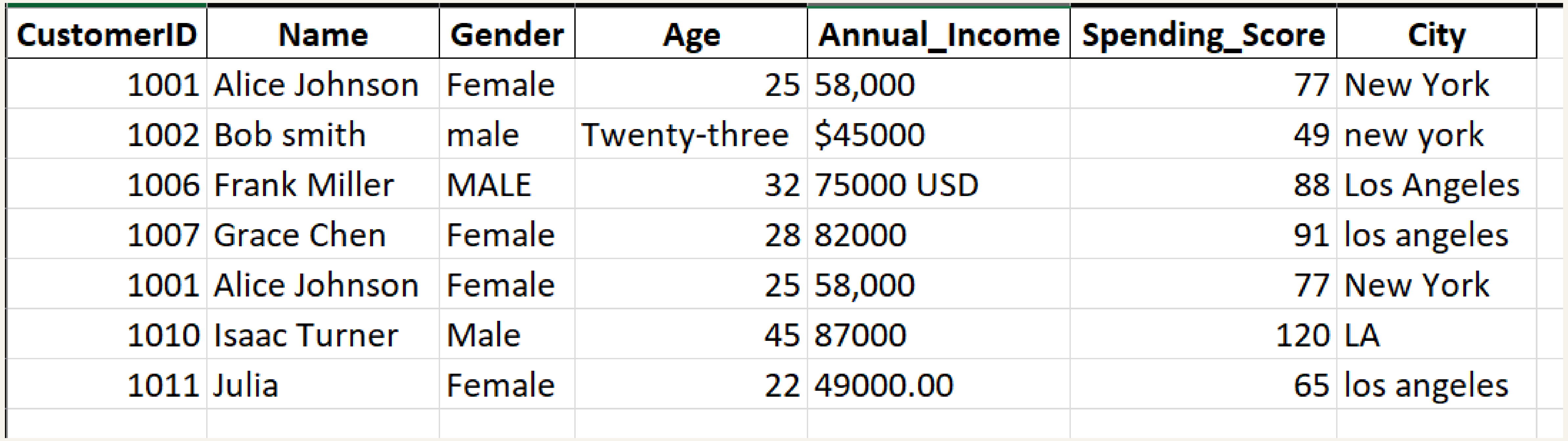
```
# Before removing duplicates
print("Rows before dropping duplicates:", df_cleaned.shape[0])

# Remove duplicates
df_no_duplicates = df_cleaned.drop_duplicates()

print("Rows after dropping duplicates:", df_no_duplicates.shape[0])
```

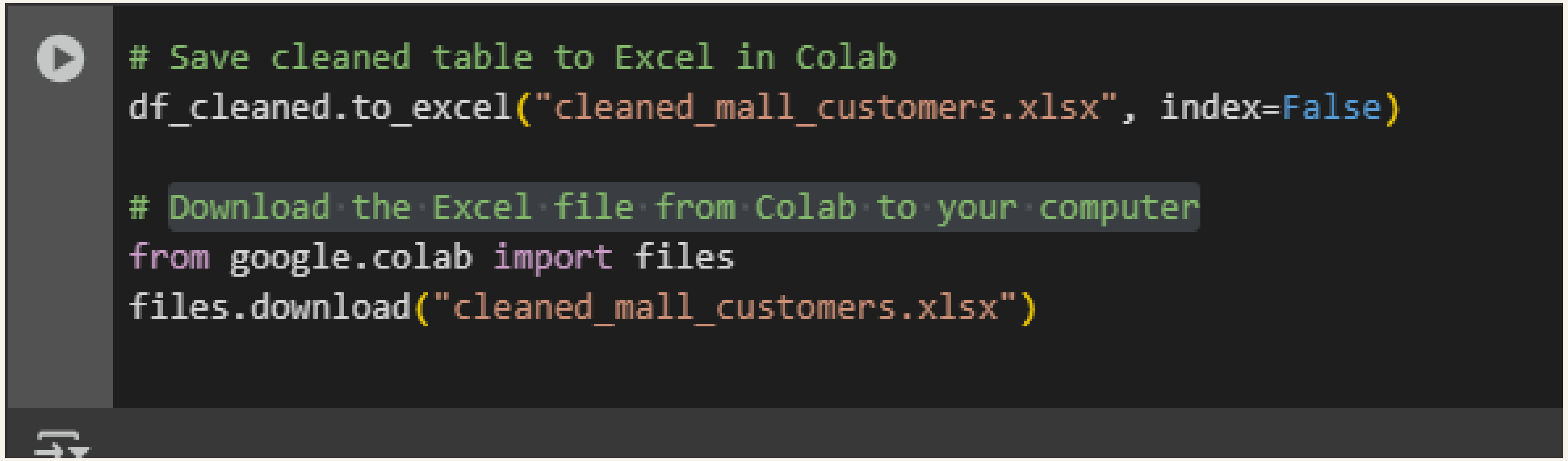
```
Rows before dropping duplicates: 7
Rows after dropping duplicates: 6
```

Result after cleaning the table



CustomerID	Name	Gender	Age	Annual_Income	Spending_Score	City
1001	Alice Johnson	Female	25	58,000	77	New York
1002	Bob smith	male	Twenty-three	$45000	49	new york
1006	Frank Miller	MALE	32	75000 USD	88	Los Angeles
1007	Grace Chen	Female	28	82000	91	los angeles
1001	Alice Johnson	Female	25	58,000	77	New York
1010	Isaac Turner	Male	45	87000	120	LA
1011	Julia	Female	22	49000.00	65	los angeles

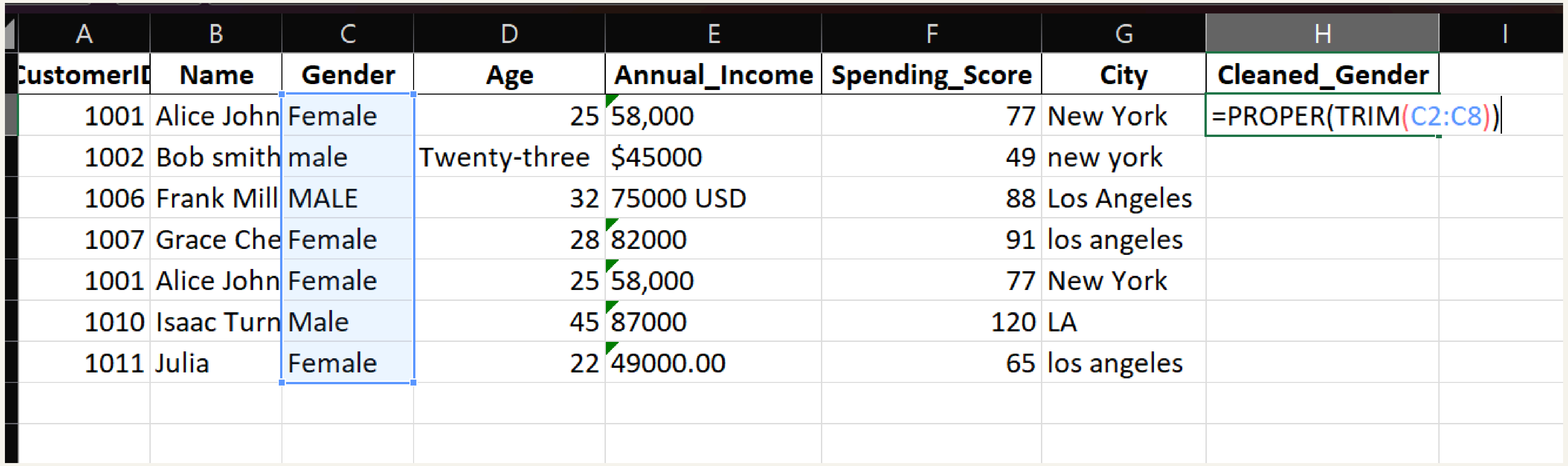
Importing file from google colab



```
# Save cleaned table to Excel in Colab
df_cleaned.to_excel("cleaned_mall_customers.xlsx", index=False)

# Download the Excel file from Colab to your computer
from google.colab import files
files.download("cleaned_mall_customers.xlsx")
```

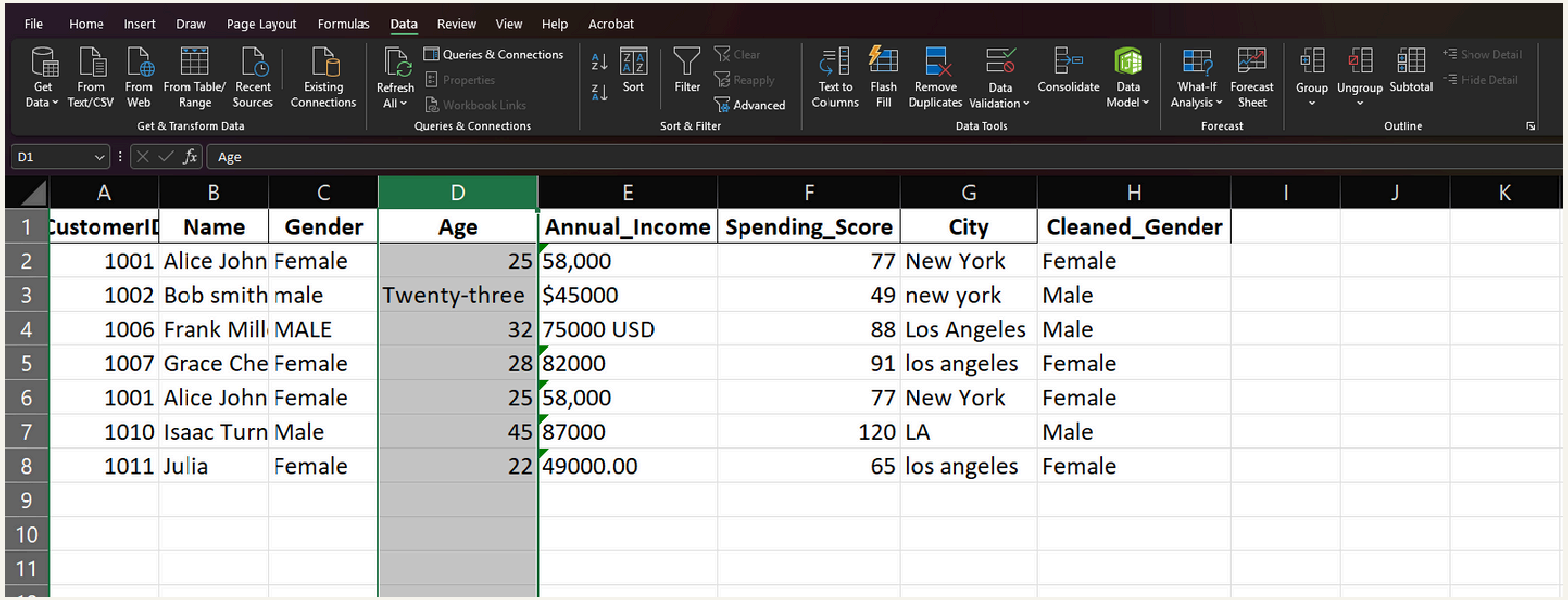
Using proper function to change the format of gender in excel



CustomerID	Name	Gender	Age	Annual_Income	Spending_Score	City	Cleaned_Gender
1001	Alice John	Female	25	58,000	77	New York	=PROPER(TRIM(C2:C8))
1002	Bob smith	male	Twenty-three	$45000	49	new york	
1006	Frank Mill	MALE	32	75000 USD	88	Los Angeles	
1007	Grace Che	Female	28	82000	91	los angeles	
1001	Alice John	Female	25	58,000	77	New York	
1010	Isaac Turn	Male	45	87000	120	LA	
1011	Julia	Female	22	49000.00	65	los angeles	

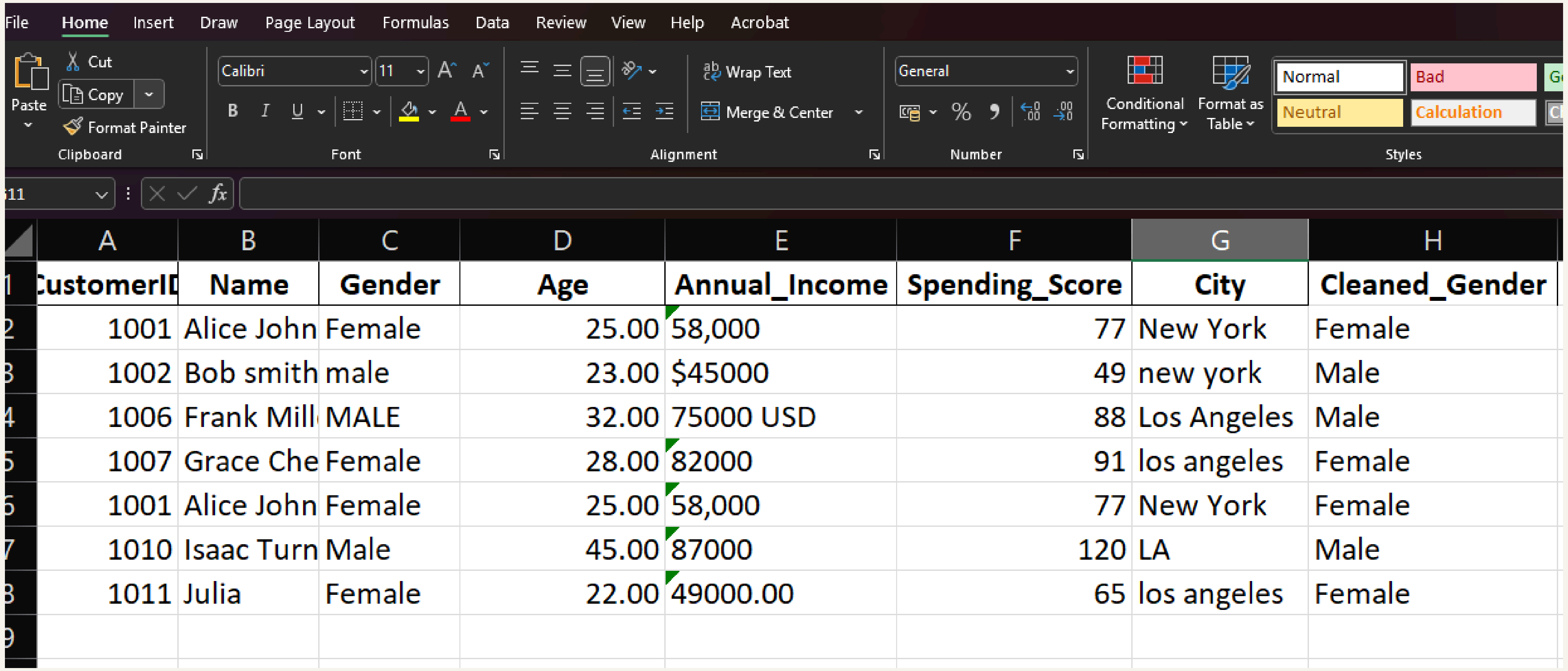
Converting age from text to column

Step 1 - Went to data
Step 2 - Selected The column of age
Step 3 - Chose text to column



	A	B	C	D	E	F	G	H	I	J	K
1	ustomerID	Name	Gender	Age	Annual_Income	Spending_Score	City	Cleaned_Gender			
2	1001	Alice John	Female	25	58,000	77	New York	Female			
3	1002	Bob smith	male	Twenty-three	$45000	49	new york	Male			
4	1006	Frank Mill	MALE	32	75000 USD	88	Los Angeles	Male			
5	1007	Grace Che	Female	28	82000	91	los angeles	Female			
6	1001	Alice John	Female	25	58,000	77	New York	Female			
7	1010	Isaac Turn	Male	45	87000	120	LA	Male			
8	1011	Julia	Female	22	49000.00	65	los angeles	Female			
9											
10											
11											

Went to home tab to change the type of number of two decimal point



	A	B	C	D	E	F	G	H
1	CustomerID	Name	Gender	Age	Annual_Income	Spending_Score	City	Cleaned_Gender
2	1001	Alice John	Female	25.00	58,000	77	New York	Female
3	1002	Bob smith	male	23.00	$45000	49	new york	Male
4	1006	Frank Mill	MALE	32.00	75000 USD	88	Los Angeles	Male
5	1007	Grace Che	Female	28.00	82000	91	los angeles	Female
6	1001	Alice John	Female	25.00	58,000	77	New York	Female
7	1010	Isaac Turn	Male	45.00	87000	120	LA	Male
8	1011	Julia	Female	22.00	49000.00	65	los angeles	Female
9								