APPLIED DATA SCIENCE

ASSIGNMENT NO : 2

TITLE : Python for Data Handling: Normalization & Standardization.

NAME: Sandip Dattatray Jadhav

ROLL NO : 82

CLASS : ECE

GITHUB LINK :

https://github.com/sandipjadhav87/Applied_Data_Science/tree/main/Assignment_2

LINK: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

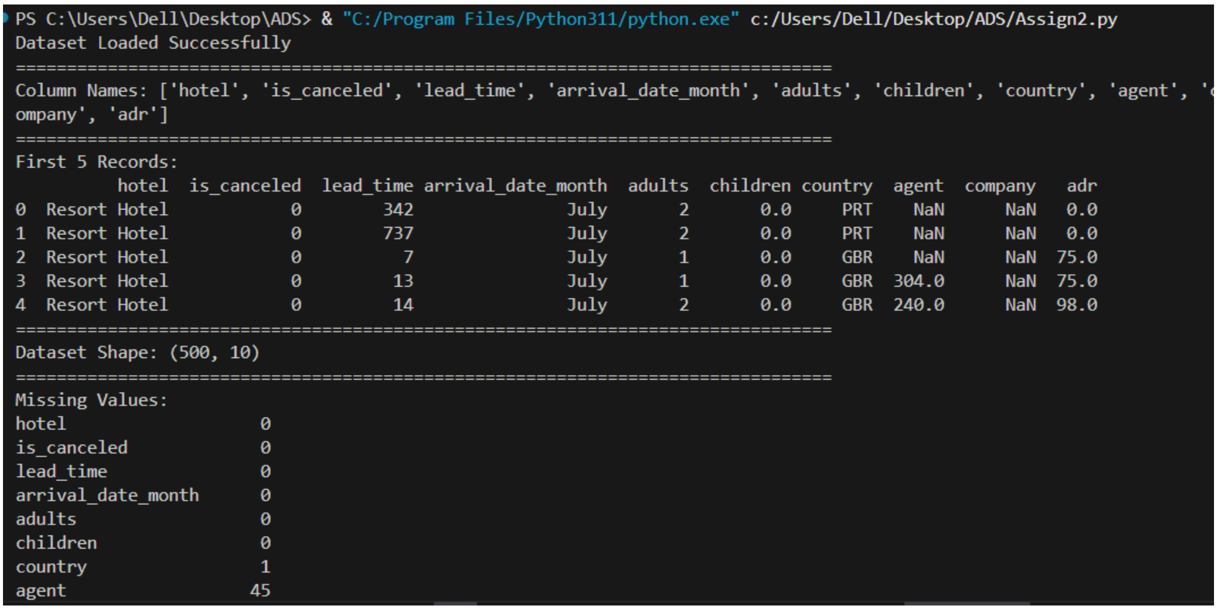
CODE:

```
# =======================================================
# APPLIED DATA SCIENCE
# Assignment 2 – Normalization & Standardization
# Dataset: Hotel Booking Demand (500 Records)
# =======================================================

# 1. Import Libraries
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler, StandardScaler

# -------------------------------------------------------
# 2. Load Dataset
# -------------------------------------------------------
df = pd.read_csv("hotel_500_records.csv")

print("Dataset Loaded Successfully")
print("=" * 80)

# -------------------------------------------------------
# 3. Clean Column Names
# -------------------------------------------------------
df.columns = df.columns.str.strip().str.lower()

print("Column Names:", df.columns.tolist())
print("=" * 80)

# -------------------------------------------------------
# 4. Basic Dataset Information
# -------------------------------------------------------
```

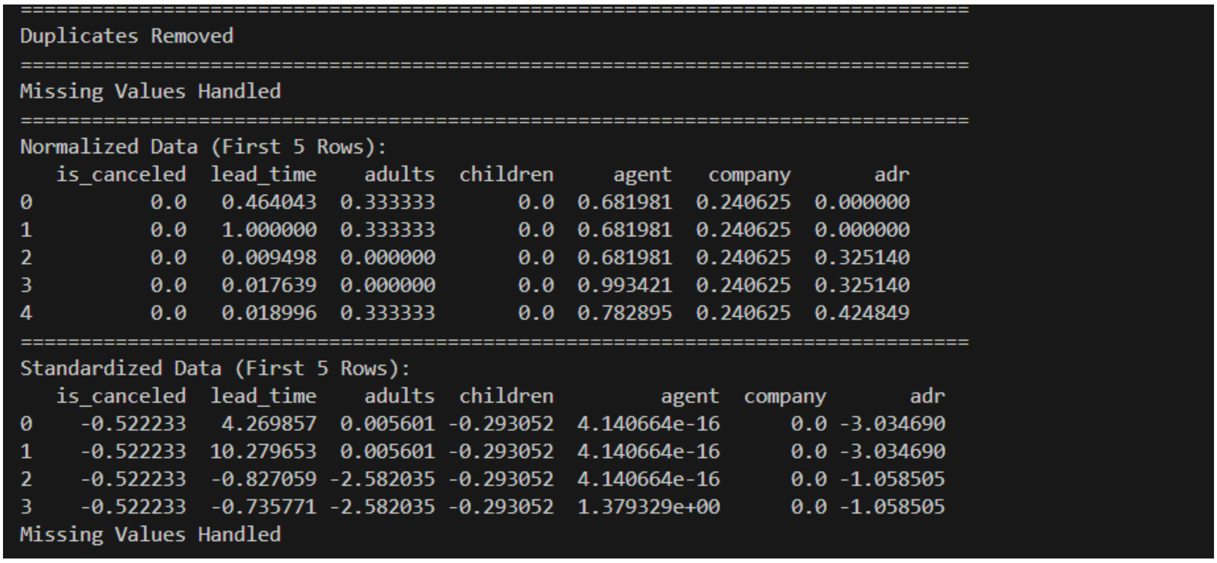
```
print("First 5 Records:")
print(df.head())
print("=" * 80)

print("Dataset Shape:", df.shape)
print("=" * 80)

print("Missing Values:")
print(df.isnull().sum())
print("=" * 80)

# ---------------------------------------------------------
# 5. Remove Duplicates
# ---------------------------------------------------------
df.drop_duplicates(inplace=True)
print("Duplicates Removed")
print("=" * 80)

# ---------------------------------------------------------
# 6. Handle Missing Values
# ---------------------------------------------------------
numerical_cols = df.select_dtypes(include=np.number).columns
df[numerical_cols] = df[numerical_cols].fillna(df[numerical_cols].mean())

cat_cols = df.select_dtypes(include=['object']).columns
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

print("Missing Values Handled")
print("=" * 80)

# ---------------------------------------------------------
# 7. Select Numerical Columns for Scaling
# ---------------------------------------------------------
numeric_data = df.select_dtypes(include=np.number)

# ---------------------------------------------------------
# 8. Normalization (Min-Max Scaling)
# ---------------------------------------------------------
minmax = MinMaxScaler()
normalized = minmax.fit_transform(numeric_data)

normalized_df = pd.DataFrame(normalized, columns=numeric_data.columns)

print("Normalized Data (First 5 Rows):")
print(normalized_df.head())
print("=" * 80)

# ---------------------------------------------------------
# 9. Standardization (Z-Score Scaling)
# ---------------------------------------------------------
```

```
standard = StandardScaler()
standardized = standard.fit_transform(numeric_data)

standardized_df = pd.DataFrame(standardized, columns=numeric_data.columns)

print("Standardized Data (First 5 Rows):")
print(standardized_df.head())

print("Data Handling Completed Successfully")
```

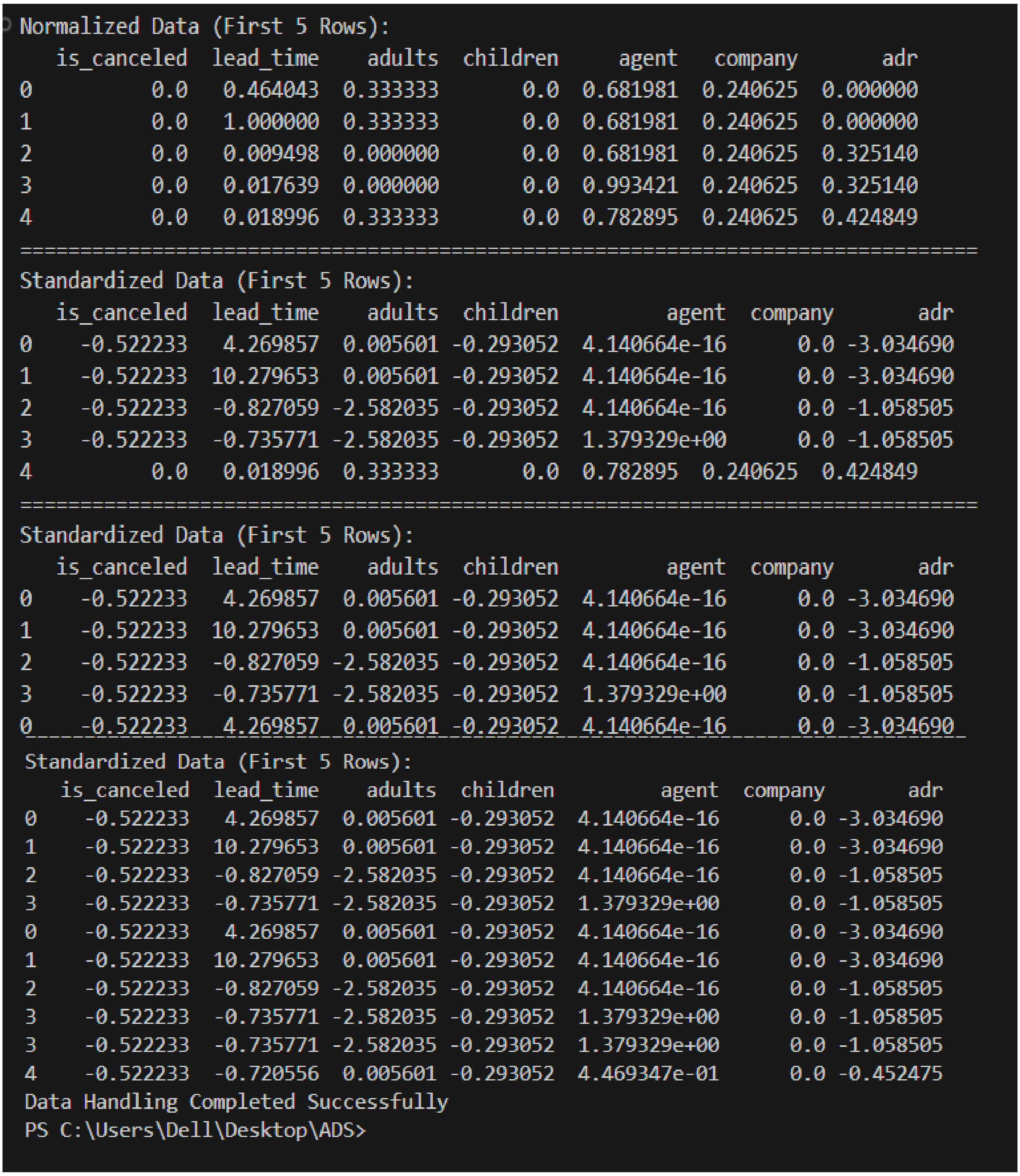
OUTPUT:



```
PS C:\Users\Dell\Desktop\ADS> & "C:/Program Files/Python311/python.exe" c:/Users/Dell/Desktop/ADS/Assign2.py
Dataset Loaded Successfully
================================================================================
Column Names: ['hotel', 'is_canceled', 'lead_time', 'arrival_date_month', 'adults', 'children', 'country', 'agent', '
ompany', 'adr']
================================================================================
First 5 Records:
          hotel  is_canceled  lead_time arrival_date_month  adults  children country  agent  company   adr
0  Resort Hotel            0        342               July       2       0.0     PRT    NaN      NaN   0.0
1  Resort Hotel            0        737               July       2       0.0     PRT    NaN      NaN   0.0
2  Resort Hotel            0          7               July       1       0.0     GBR    NaN      NaN  75.0
3  Resort Hotel            0         13               July       1       0.0     GBR  304.0      NaN  75.0
4  Resort Hotel            0         14               July       2       0.0     GBR  240.0      NaN  98.0
================================================================================
Dataset Shape: (500, 10)
================================================================================
Missing Values:
hotel                   0
is_canceled             0
lead_time               0
arrival_date_month      0
adults                  0
children                0
country                 1
agent                  45
```



```
================================================================================
Duplicates Removed
================================================================================
Missing Values Handled
================================================================================
Normalized Data (First 5 Rows):
   is_canceled  lead_time    adults  children     agent   company       adr
0          0.0   0.464043  0.333333       0.0  0.681981  0.240625  0.000000
1          0.0   1.000000  0.333333       0.0  0.681981  0.240625  0.000000
2          0.0   0.009498  0.000000       0.0  0.681981  0.240625  0.325140
3          0.0   0.017639  0.000000       0.0  0.993421  0.240625  0.325140
4          0.0   0.018996  0.333333       0.0  0.782895  0.240625  0.424849
================================================================================
Standardized Data (First 5 Rows):
   is_canceled  lead_time    adults  children         agent  company       adr
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
1    -0.522233  10.279653  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
2    -0.522233  -0.827059 -2.582035 -0.293052  4.140664e-16      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
Missing Values Handled
```



```
Normalized Data (First 5 Rows):
   is_canceled  lead_time    adults  children     agent   company       adr
0          0.0   0.464043  0.333333       0.0  0.681981  0.240625  0.000000
1          0.0   1.000000  0.333333       0.0  0.681981  0.240625  0.000000
2          0.0   0.009498  0.000000       0.0  0.681981  0.240625  0.325140
3          0.0   0.017639  0.000000       0.0  0.993421  0.240625  0.325140
4          0.0   0.018996  0.333333       0.0  0.782895  0.240625  0.424849
================================================================================
Standardized Data (First 5 Rows):
   is_canceled  lead_time    adults  children         agent  company       adr
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
1    -0.522233  10.279653  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
2    -0.522233  -0.827059 -2.582035 -0.293052  4.140664e-16      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
4          0.0   0.018996  0.333333       0.0  0.782895  0.240625  0.424849
================================================================================
Standardized Data (First 5 Rows):
   is_canceled  lead_time    adults  children         agent  company       adr
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
1    -0.522233  10.279653  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
2    -0.522233  -0.827059 -2.582035 -0.293052  4.140664e-16      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
Standardized Data (First 5 Rows):
   is_canceled  lead_time    adults  children         agent  company       adr
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
1    -0.522233  10.279653  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
2    -0.522233  -0.827059 -2.582035 -0.293052  4.140664e-16      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
0    -0.522233   4.269857  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
1    -0.522233  10.279653  0.005601 -0.293052  4.140664e-16      0.0 -3.034690
2    -0.522233  -0.827059 -2.582035 -0.293052  4.140664e-16      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
3    -0.522233  -0.735771 -2.582035 -0.293052  1.379329e+00      0.0 -1.058505
4    -0.522233  -0.720556  0.005601 -0.293052  4.469347e-01      0.0 -0.452475
Data Handling Completed Successfully
PS C:\Users\Dell\Desktop\ADS>
```

INTERPRETATION:

- ## Interpretation of Normalization
- Normalization scaled housing features like price, area and bedrooms into a range between 0 and 1.
  All attributes now have equal scale regardless of original units.
  No negative values were produced and data distribution remained unchanged.

- ## Interpretation of Standardization
- Standardization transformed housing data so that mean became 0 and standard deviation became 1.
  Values above average became positive and below average became negative.
  This helps compare how far each house feature lies from the average.