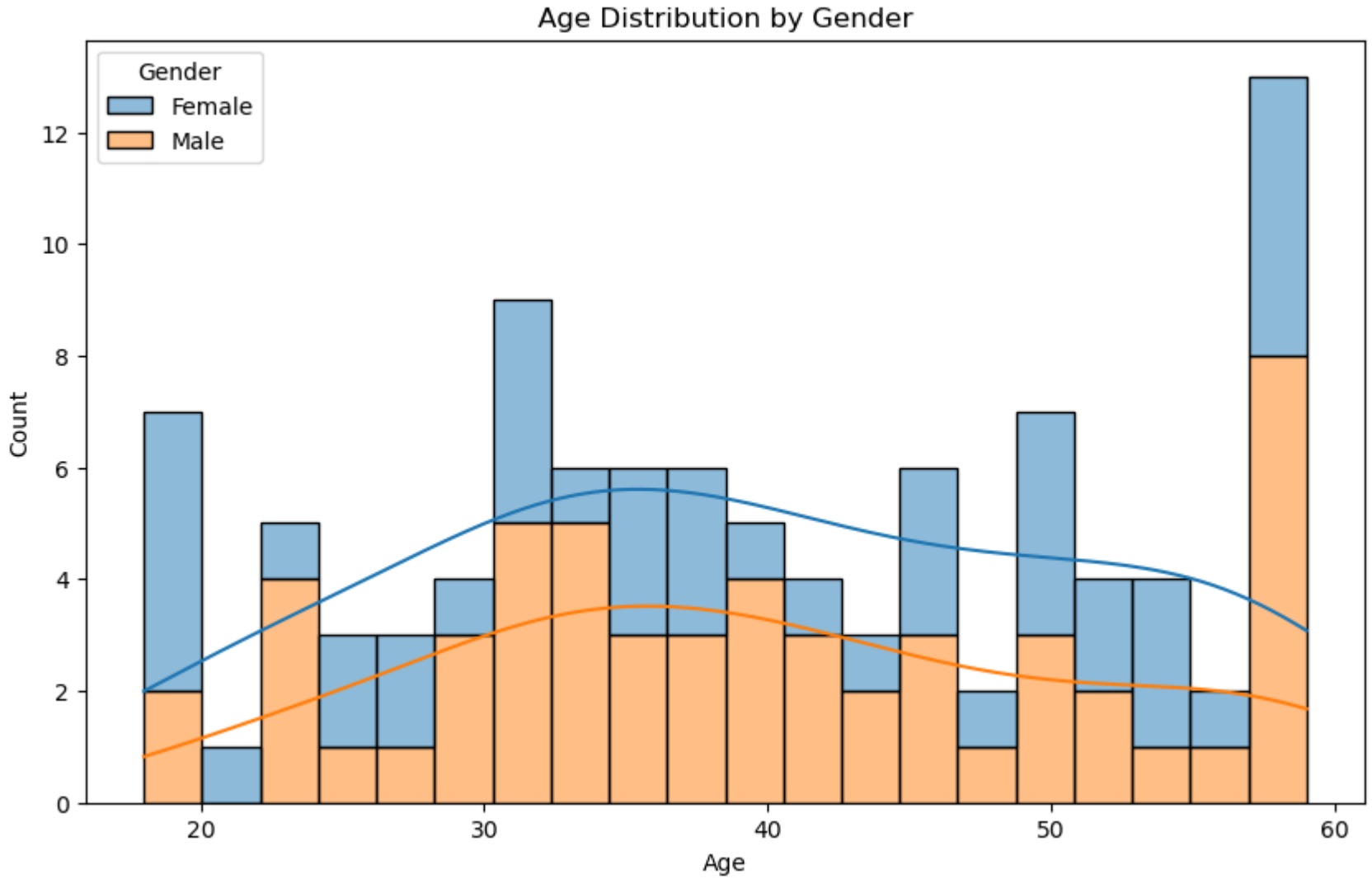
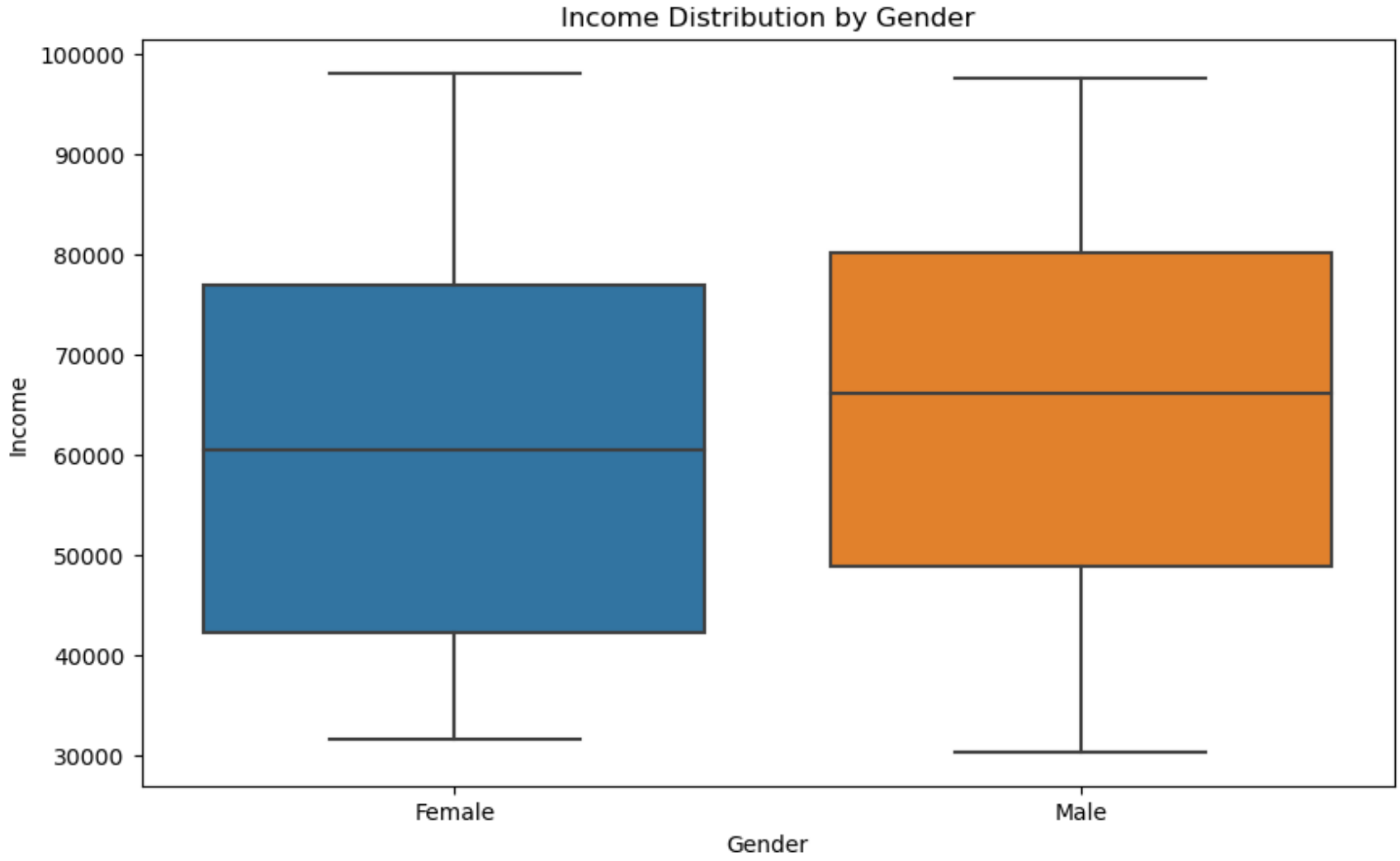
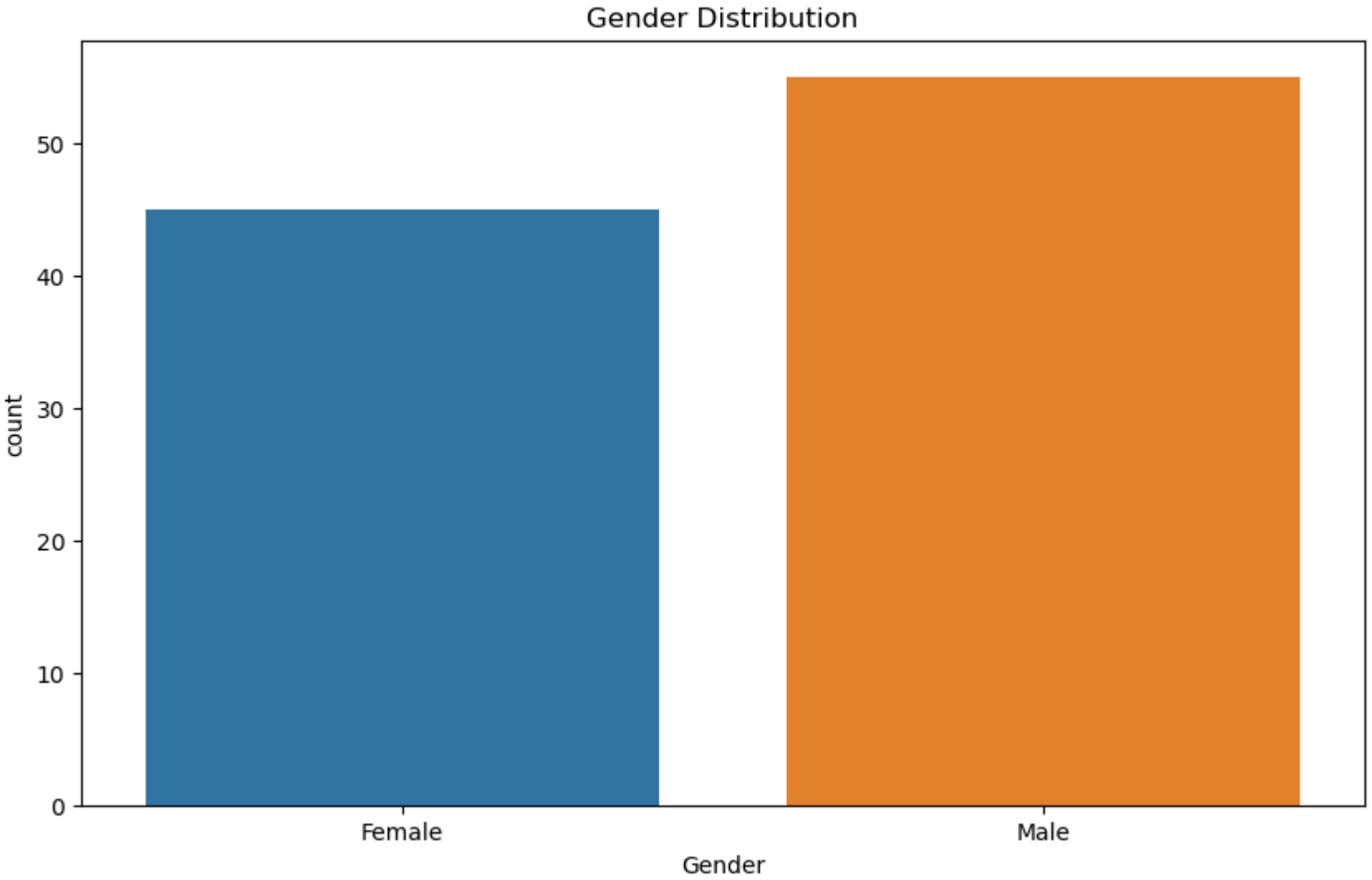
In [3]:
```
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
import matplotlib.pyplot as plt
import seaborn as sns

class UserProfilesReport:
    def __init__(self, spark_session, data_path):
        self.spark = spark_session
        self.data_path = data_path

    def load_data(self):
        return self.spark.read.csv(self.data_path, header=True, inferSchema=True)

    def preprocess_data(self, data):
        # Handle null values for numerical attribute 'Age'
        data = data.na.fill({'Age': data.select(F.mean('Age')).collect()[0][0]})
        return data

    def describe_age(self, data):
        # Profile numerical attribute 'Age'
        data.select('Age').describe().show()

    def find_outliers(self, data):
        # Explain outliers for numerical attribute 'Age'
        outliers = data.filter(F.col('Age') > 100)
        print("Outliers in Age:")
        outliers.show()

    def visualize_data(self, data):
        # Print tabular data
        print("Tabular Data:")
        data.show()

        # Plot Age Distribution by Gender
        plt.figure(figsize=(10, 6))
        sns.histplot(data=data.toPandas(), x='Age', hue='Gender', bins=20, multiple='stack', kde=True)
        plt.title('Age Distribution by Gender')
        plt.show()

        # Plot Income Distribution by Gender
        plt.figure(figsize=(10, 6))
        sns.boxplot(x='Gender', y='Income', data=data.toPandas())
        plt.title('Income Distribution by Gender')
```

```
        plt.show()

        # Plot Gender Distribution
        plt.figure(figsize=(10, 6))
        sns.countplot(x='Gender', data=data.toPandas())
        plt.title('Gender Distribution')
        plt.show()

# Example of usage
spark = SparkSession.builder.appName("UserProfilesJob").getOrCreate()
user_profiles_report = UserProfilesReport(spark, '/home/raja/Documents/MS-DATA-SCIENCE/Fall-2023/Big-Data-Pro
user_profiles_data = user_profiles_report.load_data()
user_profiles_data = user_profiles_report.preprocess_data(user_profiles_data)
user_profiles_report.describe_age(user_profiles_data)
user_profiles_report.find_outliers(user_profiles_data)
user_profiles_report.visualize_data(user_profiles_data)
```

```
+-------+------------------+
|summary|               Age|
+-------+------------------+
|  count|               100|
|   mean|             39.66|
| stddev|12.032128706597232|
|    min|                18|
|    max|                59|
+-------+------------------+

Outliers in Age:
+-------+----+---+------+------+-------+--------+
|User_ID|Name|Age|Gender|Income|Contact|Location|
+-------+----+---+------+------+-------+--------+
+-------+----+---+------+------+-------+--------+

Tabular Data:
+-------+-----------------+---+------+------+--------------------+--------------------+
|User_ID|             Name|Age|Gender|Income|             Contact|            Location|
+-------+-----------------+---+------+------+--------------------+--------------------+
|      1|    Heather Burke| 35|Female| 68059|        347-414-3911|61.389715, -7.198571|
|      2|   Margaret Brown| 40|  Male| 73613|     +1-632-865-1129|-61.234150, -9.52...|
|      3|     Pamela Crane| 53|  Male| 75572|  816.752.4431x56859|-68.287243, -135....|
|      4| Erika Williamson| 36|  Male| 63795|     +1-652-669-8271|MDR37A, Basni, Na...|
|      5|    Lori Mcintosh| 38|  Male| 31765|  777-357-5255x67102|80.204909, -172.1...|
|      6|    Desiree Logan| 36|  Male| 51571|     +1-440-563-7070|-10.650334, 13.40...|
|      7| Steven Maldonado| 32|Female| 44402|001-775-236-0974x...|-71.985498, -67.2...|
|      8|Joshua Strickland| 31|Female| 47287|+1-307-638-0387x1...|87.995936, 20.762031|
|      9|      John Hansen| 54|Female| 61199|       (803)636-5015|-83.582661, 175.9...|
|     10|   Nicholas Young| 54|Female| 65067|   (629)707-1373x986|-31.759288, 105.8...|
|     11|     Crystal Rice| 40|  Male| 85409|        714-835-1574|76.212865, 31.831365|
|     12|       Drew Hicks| 58|  Male| 66058| (241)872-3581x43080|Municipio de Fiam...|
|     13|     April Wright| 46|  Male| 47672|          9799875151|70.867290, 162.96...|
|     14|    Lucas Ballard| 29|  Male| 88190|        288-705-8340|  Tagant, Mauritania|
|     15|       Laura Reed| 40|Female| 32865|   411.801.0944x4420|-42.352781, -176....|
|     16|     Sean Brennan| 51|  Male| 69204|    769.608.5873x453|15.157090, -102.0...|
|     17|   Brittany Jones| 33|  Male| 97613|     +1-492-733-0050|-40.271602, -41.9...|
|     18|    Lori Robinson| 31|  Male| 84333|          7428762933|Qikiqtaaluk Regio...|
|     19|    Paige Johnson| 19|Female| 64488|+1-245-945-8598x3...|87.115777, 174.23...|
|     20|     Susan Martin| 28|  Male| 53571|+1-277-764-3770x2...|-77.714473, 126.1...|
+-------+-----------------+---+------+------+--------------------+--------------------+
only showing top 20 rows
```


Age Distribution by Gender
Gender
Female
Male
Count
Age
0
2
4
6
8
10
12
20
30
40
50
60


Income Distribution by Gender
100000
90000
80000
70000
60000
50000
40000
30000
Income
Female
Male
Gender


Gender Distribution
count
0
10
20
30
40
50
Female
Male
Gender