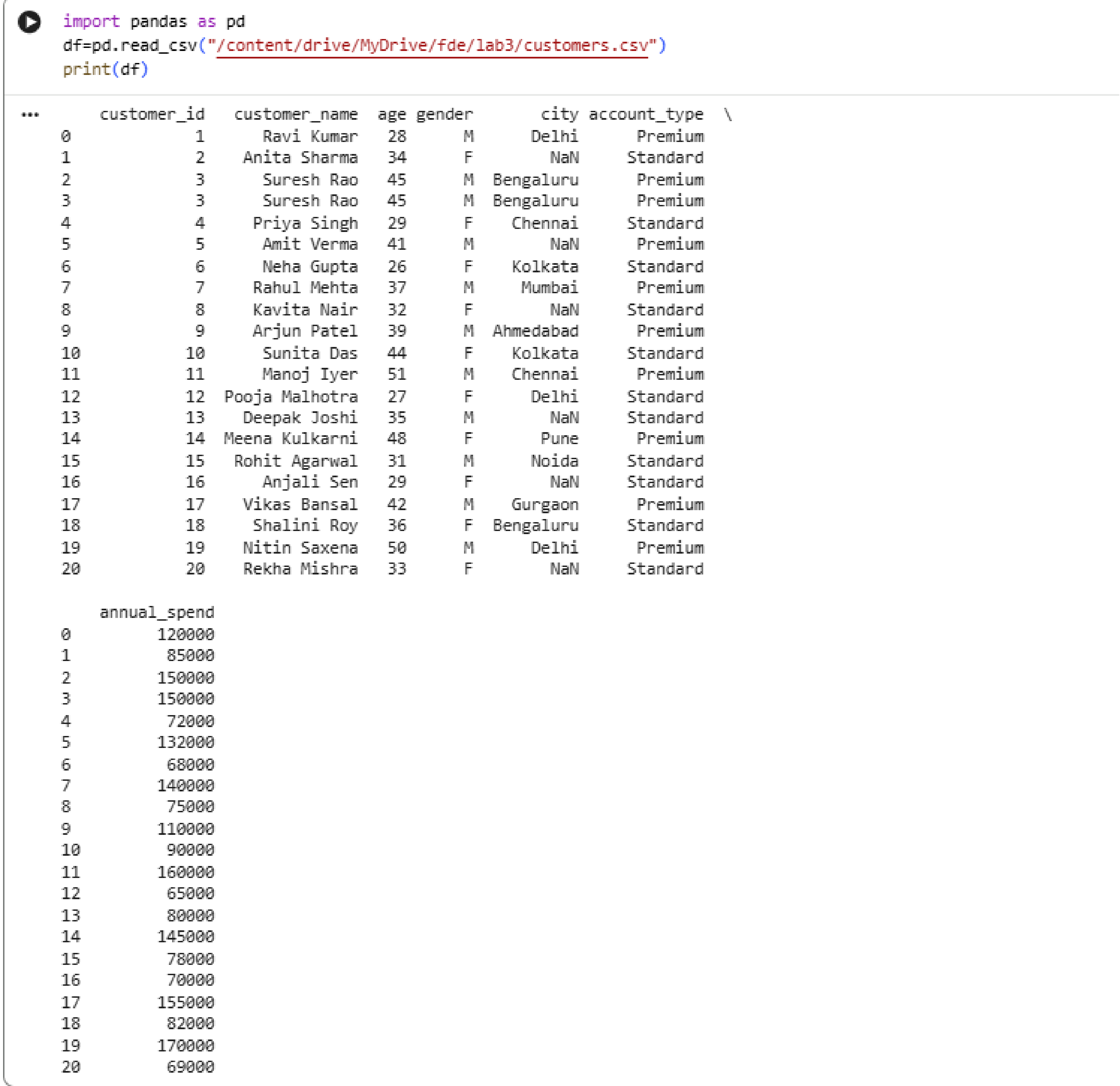
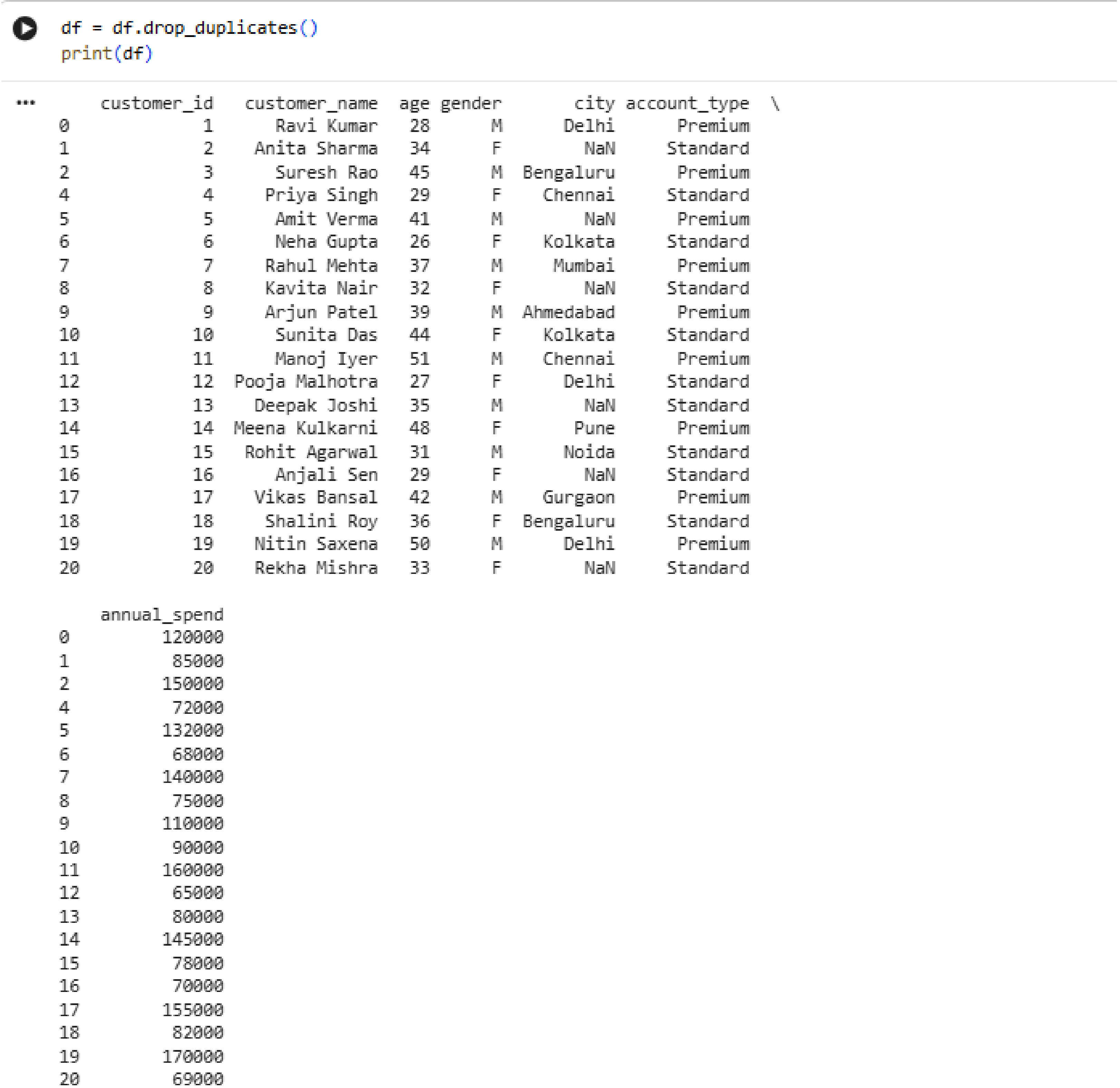


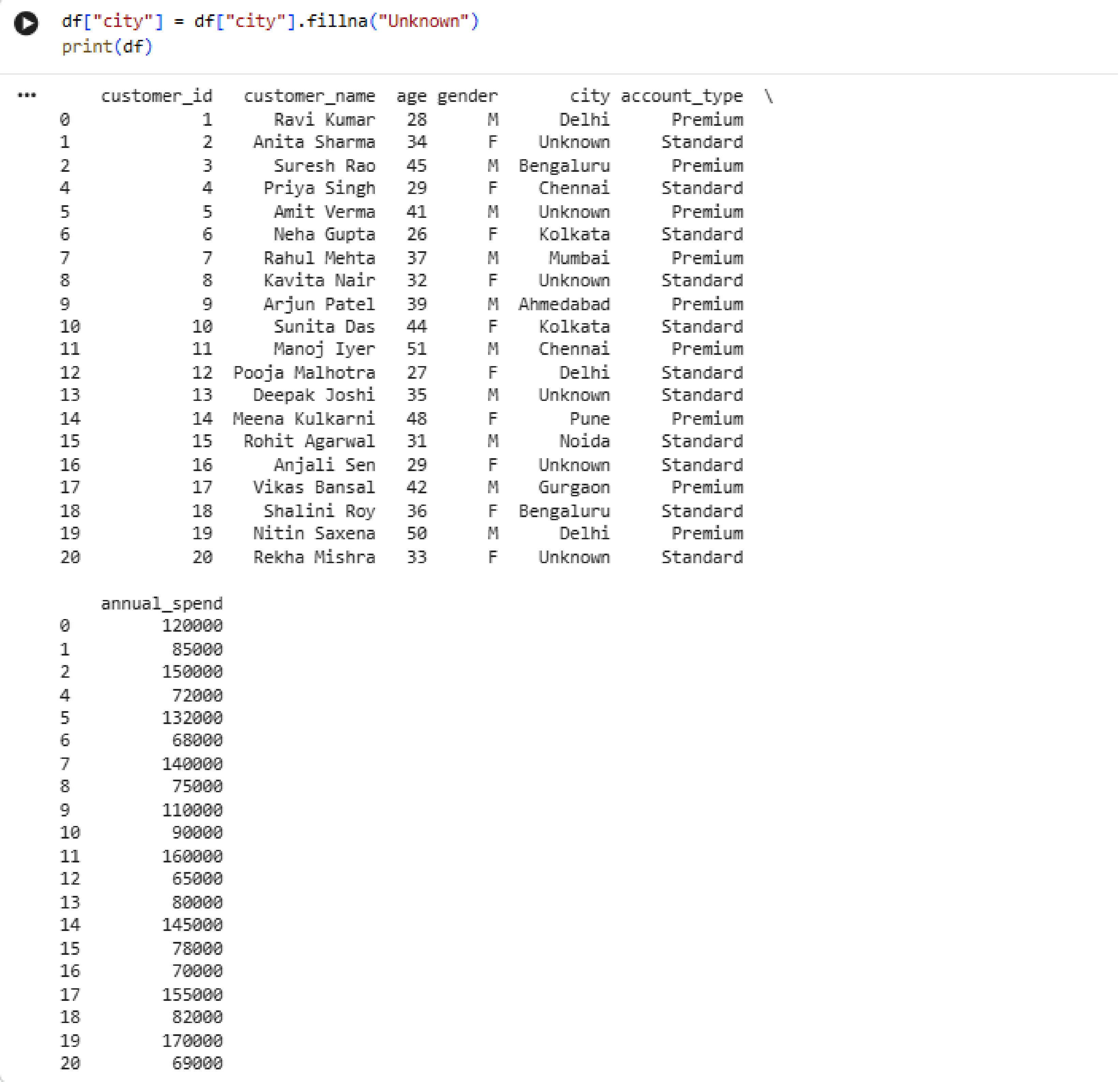
```
import pandas as pd
df=pd.read_csv("/content/drive/MyDrive/fde/lab3/customers.csv")
print(df)
```

```
    customer_id   customer_name  age gender       city account_type  \
0             1      Ravi Kumar   28      M      Delhi      Premium   
1             2    Anita Sharma   34      F        NaN     Standard   
2             3      Suresh Rao   45      M  Bengaluru      Premium   
3             3      Suresh Rao   45      M  Bengaluru      Premium   
4             4     Priya Singh   29      F    Chennai     Standard   
5             5      Amit Verma   41      M        NaN      Premium   
6             6      Neha Gupta   26      F    Kolkata     Standard   
7             7     Rahul Mehta   37      M     Mumbai      Premium   
8             8     Kavita Nair   32      F        NaN     Standard   
9             9     Arjun Patel   39      M  Ahmedabad      Premium   
10           10      Sunita Das   44      F    Kolkata     Standard   
11           11      Manoj Iyer   51      M    Chennai      Premium   
12           12  Pooja Malhotra   27      F      Delhi     Standard   
13           13    Deepak Joshi   35      M        NaN     Standard   
14           14  Meena Kulkarni   48      F       Pune      Premium   
15           15   Rohit Agarwal   31      M      Noida     Standard   
16           16      Anjali Sen   29      F        NaN     Standard   
17           17    Vikas Bansal   42      M    Gurgaon      Premium   
18           18     Shalini Roy   36      F  Bengaluru     Standard   
19           19    Nitin Saxena   50      M      Delhi      Premium   
20           20    Rekha Mishra   33      F        NaN     Standard   

    annual_spend  
0         120000  
1          85000  
2         150000  
3         150000  
4          72000  
5         132000  
6          68000  
7         140000  
8          75000  
9         110000  
10         90000  
11        160000  
12         65000  
13         80000  
14        145000  
15         78000  
16         70000  
17        155000  
18         82000  
19        170000  
20         69000  
```



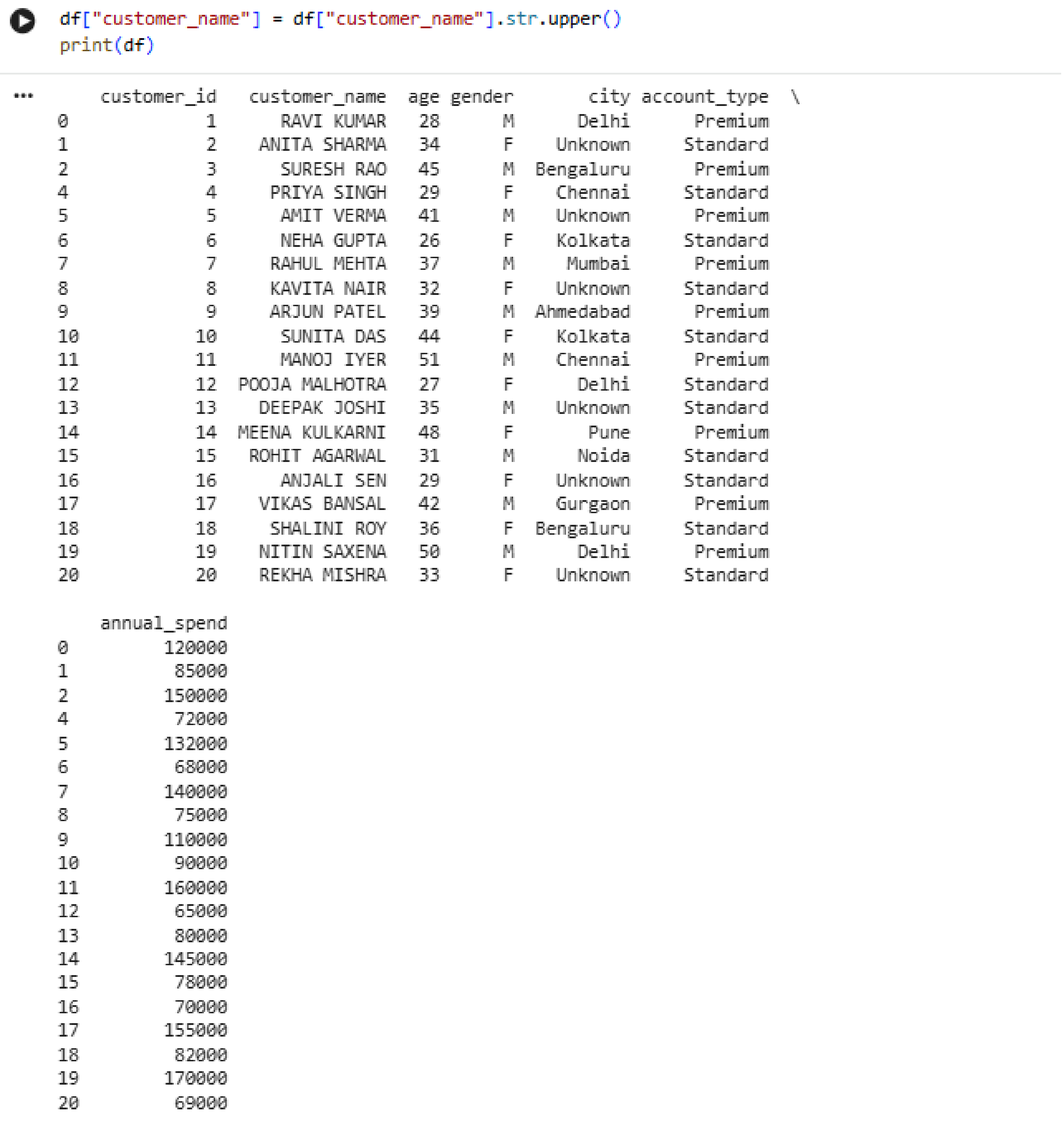
```
df = df.drop_duplicates()
print(df)
```

```
    customer_id   customer_name  age gender       city account_type  \
0             1      Ravi Kumar   28      M      Delhi      Premium
1             2    Anita Sharma   34      F        NaN     Standard
2             3      Suresh Rao   45      M  Bengaluru      Premium
4             4     Priya Singh   29      F    Chennai     Standard
5             5      Amit Verma   41      M        NaN      Premium
6             6      Neha Gupta   26      F    Kolkata     Standard
7             7     Rahul Mehta   37      M     Mumbai      Premium
8             8     Kavita Nair   32      F        NaN     Standard
9             9     Arjun Patel   39      M  Ahmedabad      Premium
10           10      Sunita Das   44      F    Kolkata     Standard
11           11      Manoj Iyer   51      M    Chennai      Premium
12           12  Pooja Malhotra   27      F      Delhi     Standard
13           13    Deepak Joshi   35      M        NaN     Standard
14           14  Meena Kulkarni   48      F       Pune      Premium
15           15   Rohit Agarwal   31      M      Noida     Standard
16           16      Anjali Sen   29      F        NaN     Standard
17           17    Vikas Bansal   42      M    Gurgaon      Premium
18           18     Shalini Roy   36      F  Bengaluru     Standard
19           19    Nitin Saxena   50      M      Delhi      Premium
20           20    Rekha Mishra   33      F        NaN     Standard

    annual_spend
0         120000
1          85000
2         150000
4          72000
5         132000
6          68000
7         140000
8          75000
9         110000
10         90000
11        160000
12         65000
13         80000
14        145000
15         78000
16         70000
17        155000
18         82000
19        170000
20         69000
```



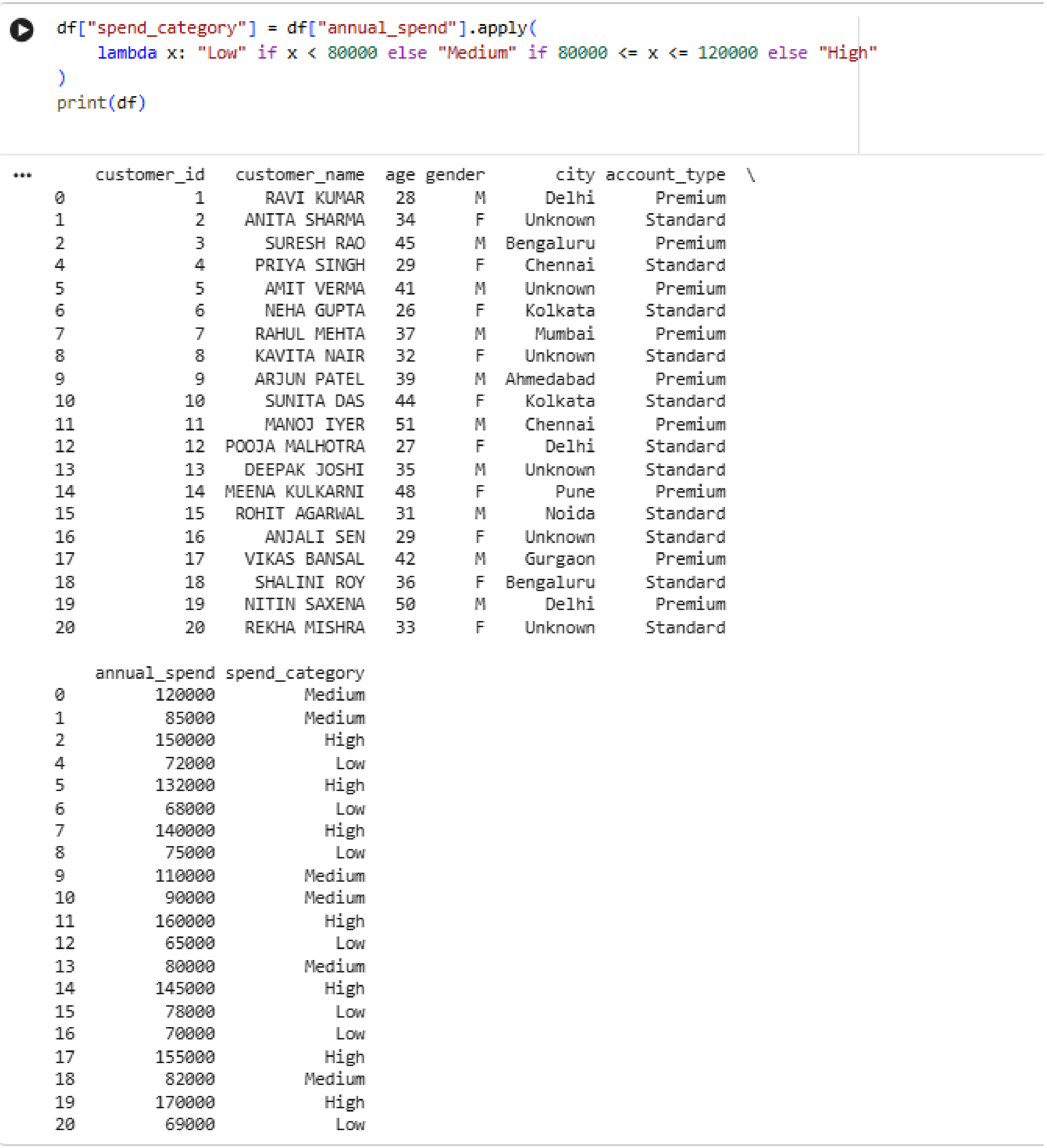
```
df["city"] = df["city"].fillna("Unknown")
print(df)
```

```
    customer_id   customer_name  age gender       city account_type  \
0             1      Ravi Kumar   28      M      Delhi      Premium   
1             2    Anita Sharma   34      F    Unknown     Standard   
2             3      Suresh Rao   45      M  Bengaluru      Premium   
4             4     Priya Singh   29      F    Chennai     Standard   
5             5      Amit Verma   41      M    Unknown      Premium   
6             6      Neha Gupta   26      F    Kolkata     Standard   
7             7     Rahul Mehta   37      M     Mumbai      Premium   
8             8     Kavita Nair   32      F    Unknown     Standard   
9             9     Arjun Patel   39      M  Ahmedabad      Premium   
10           10      Sunita Das   44      F    Kolkata     Standard   
11           11      Manoj Iyer   51      M    Chennai      Premium   
12           12  Pooja Malhotra   27      F      Delhi     Standard   
13           13    Deepak Joshi   35      M    Unknown     Standard   
14           14  Meena Kulkarni   48      F       Pune      Premium   
15           15   Rohit Agarwal   31      M      Noida     Standard   
16           16      Anjali Sen   29      F    Unknown     Standard   
17           17    Vikas Bansal   42      M    Gurgaon      Premium   
18           18     Shalini Roy   36      F  Bengaluru     Standard   
19           19    Nitin Saxena   50      M      Delhi      Premium   
20           20    Rekha Mishra   33      F    Unknown     Standard   

    annual_spend  
0         120000  
1          85000  
2         150000  
4          72000  
5         132000  
6          68000  
7         140000  
8          75000  
9         110000  
10         90000  
11        160000  
12         65000  
13         80000  
14        145000  
15         78000  
16         70000  
17        155000  
18         82000  
19        170000  
20         69000  
```



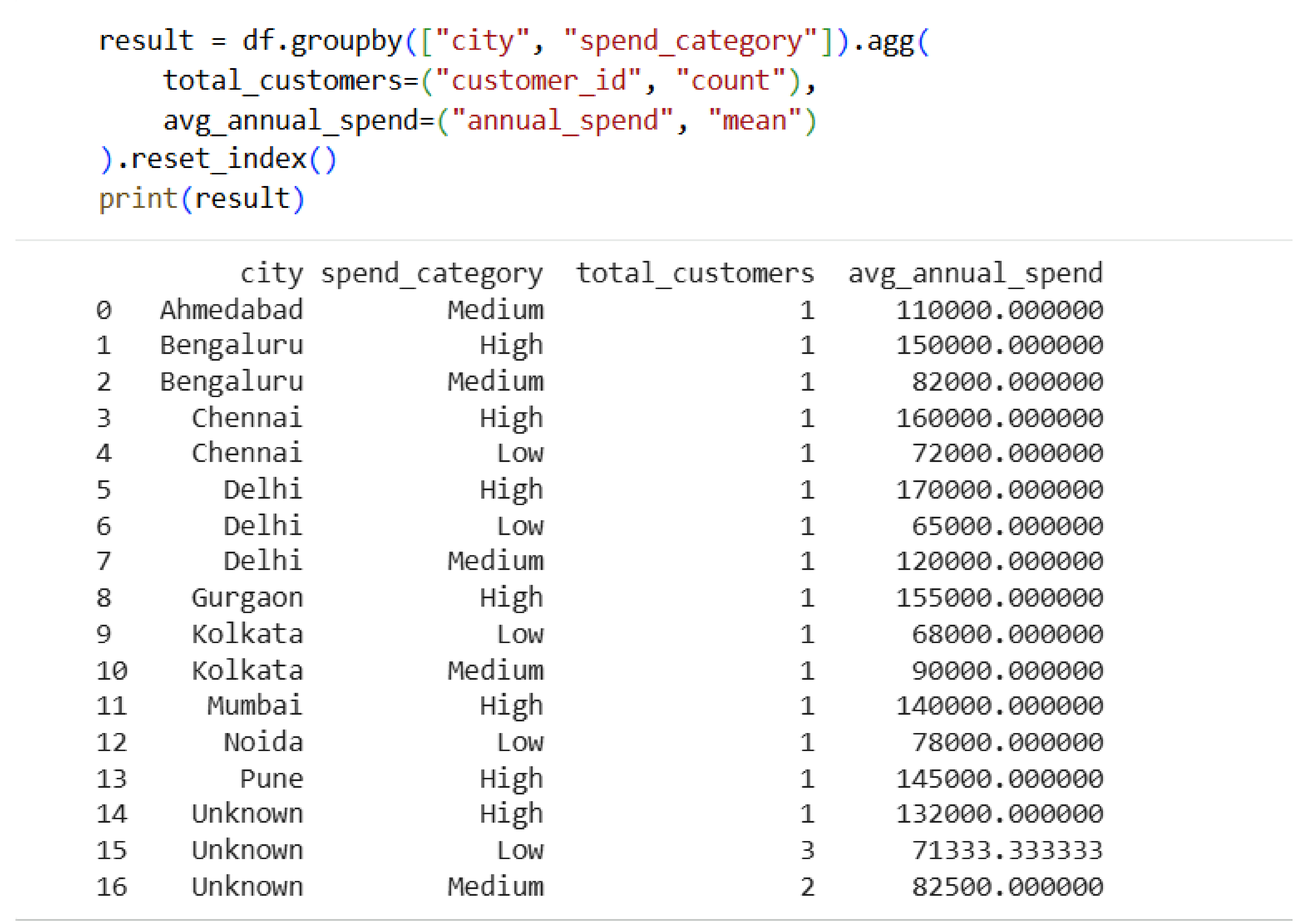
```
df["customer_name"] = df["customer_name"].str.upper()
print(df)
```

```
    customer_id   customer_name  age gender       city account_type  \
0             1      RAVI KUMAR   28      M      Delhi      Premium   
1             2    ANITA SHARMA   34      F    Unknown     Standard   
2             3      SURESH RAO   45      M  Bengaluru      Premium   
4             4     PRIYA SINGH   29      F    Chennai     Standard   
5             5      AMIT VERMA   41      M    Unknown      Premium   
6             6      NEHA GUPTA   26      F    Kolkata     Standard   
7             7     RAHUL MEHTA   37      M     Mumbai      Premium   
8             8     KAVITA NAIR   32      F    Unknown     Standard   
9             9     ARJUN PATEL   39      M  Ahmedabad      Premium   
10           10      SUNITA DAS   44      F    Kolkata     Standard   
11           11      MANOJ IYER   51      M    Chennai      Premium   
12           12  POOJA MALHOTRA   27      F      Delhi     Standard   
13           13    DEEPAK JOSHI   35      M    Unknown     Standard   
14           14  MEENA KULKARNI   48      F       Pune      Premium   
15           15   ROHIT AGARWAL   31      M      Noida     Standard   
16           16      ANJALI SEN   29      F    Unknown     Standard   
17           17    VIKAS BANSAL   42      M    Gurgaon      Premium   
18           18     SHALINI ROY   36      F  Bengaluru     Standard   
19           19    NITIN SAXENA   50      M      Delhi      Premium   
20           20    REKHA MISHRA   33      F    Unknown     Standard   

    annual_spend  
0         120000  
1          85000  
2         150000  
4          72000  
5         132000  
6          68000  
7         140000  
8          75000  
9         110000  
10         90000  
11        160000  
12         65000  
13         80000  
14        145000  
15         78000  
16         70000  
17        155000  
18         82000  
19        170000  
20         69000  
```



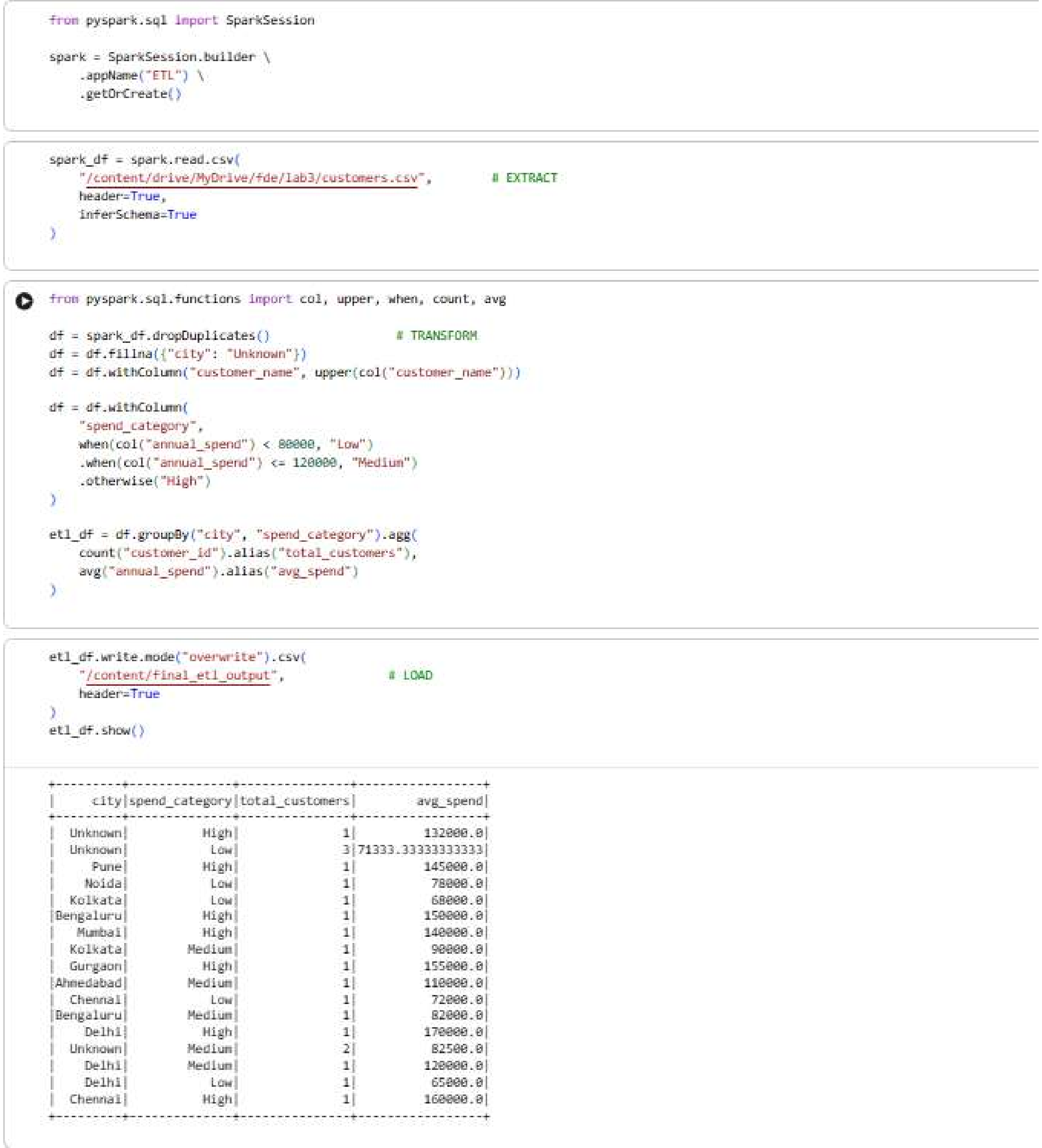
```
df["spend_category"] = df["annual_spend"].apply(
    lambda x: "Low" if x < 80000 else "Medium" if 80000 <= x <= 120000 else "High"
)
print(df)
```

```
    customer_id   customer_name  age gender       city account_type  \
0             1      RAVI KUMAR   28      M      Delhi      Premium
1             2    ANITA SHARMA   34      F    Unknown     Standard
2             3      SURESH RAO   45      M  Bengaluru      Premium
4             4     PRIYA SINGH   29      F    Chennai     Standard
5             5      AMIT VERMA   41      M    Unknown      Premium
6             6      NEHA GUPTA   26      F    Kolkata     Standard
7             7     RAHUL MEHTA   37      M     Mumbai      Premium
8             8     KAVITA NAIR   32      F    Unknown     Standard
9             9     ARJUN PATEL   39      M  Ahmedabad      Premium
10           10      SUNITA DAS   44      F    Kolkata     Standard
11           11      MANOJ IYER   51      M    Chennai      Premium
12           12  POOJA MALHOTRA   27      F      Delhi     Standard
13           13    DEEPAK JOSHI   35      M    Unknown     Standard
14           14  MEENA KULKARNI   48      F       Pune      Premium
15           15   ROHIT AGARWAL   31      M      Noida     Standard
16           16      ANJALI SEN   29      F    Unknown     Standard
17           17    VIKAS BANSAL   42      M    Gurgaon      Premium
18           18     SHALINI ROY   36      F  Bengaluru     Standard
19           19    NITIN SAXENA   50      M      Delhi      Premium
20           20    REKHA MISHRA   33      F    Unknown     Standard

    annual_spend spend_category
0         120000         Medium
1          85000         Medium
2         150000           High
4          72000            Low
5         132000           High
6          68000            Low
7         140000           High
8          75000            Low
9         110000         Medium
10         90000         Medium
11        160000           High
12         65000            Low
13         80000         Medium
14        145000           High
15         78000            Low
16         70000            Low
17        155000           High
18         82000         Medium
19        170000           High
20         69000            Low
```



```
result = df.groupby(["city", "spend_category"]).agg(
    total_customers=("customer_id", "count"),
    avg_annual_spend=("annual_spend", "mean")
).reset_index()
print(result)
```

```
         city spend_category  total_customers  avg_annual_spend
0   Ahmedabad         Medium                1     110000.000000
1   Bengaluru           High                1     150000.000000
2   Bengaluru         Medium                1      82000.000000
3     Chennai           High                1     160000.000000
4     Chennai            Low                1      72000.000000
5       Delhi           High                1     170000.000000
6       Delhi            Low                1      65000.000000
7       Delhi         Medium                1     120000.000000
8     Gurgaon           High                1     155000.000000
9     Kolkata            Low                1      68000.000000
10    Kolkata         Medium                1      90000.000000
11     Mumbai           High                1     140000.000000
12      Noida            Low                1      78000.000000
13       Pune           High                1     145000.000000
14    Unknown           High                1     132000.000000
15    Unknown            Low                3      71333.333333
16    Unknown         Medium                2      82500.000000
```



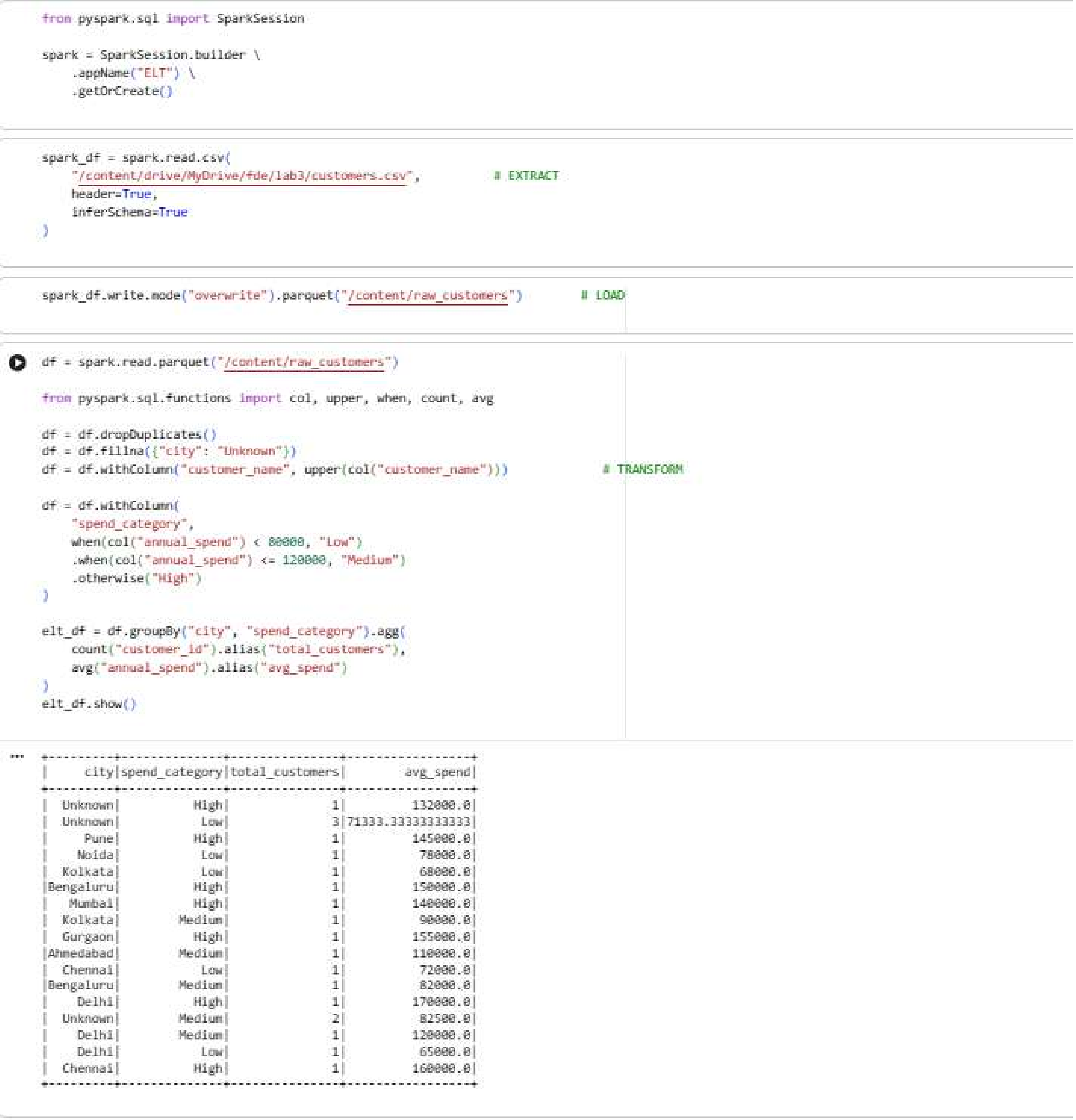
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("ETL") \
    .getOrCreate()
```

```
spark_df = spark.read.csv(
    "/content/drive/MyDrive/fde/lab3/customers.csv",        # EXTRACT
    header=True,
    inferSchema=True
)
```

```
from pyspark.sql.functions import col, upper, when, count, avg

df = spark_df.dropDuplicates()                 # TRANSFORM
df = df.fillna({"city": "Unknown"})
df = df.withColumn("customer_name", upper(col("customer_name")))

df = df.withColumn(
    "spend_category",
    when(col("annual_spend") < 80000, "Low")
    .when(col("annual_spend") <= 120000, "Medium")
    .otherwise("High")
)

etl_df = df.groupBy("city", "spend_category").agg(
    count("customer_id").alias("total_customers"),
    avg("annual_spend").alias("avg_spend")
)
```

```
etl_df.write.mode("overwrite").csv(
    "/content/final_etl_output",              # LOAD
    header=True
)
etl_df.show()
```

```
+---------+--------------+---------------+-----------------+
|     city|spend_category|total_customers|        avg_spend|
+---------+--------------+---------------+-----------------+
|  Unknown|          High|              1|         132000.0|
|  Unknown|           Low|              3|71333.33333333333|
|     Pune|          High|              1|         145000.0|
|    Noida|           Low|              1|          78000.0|
|  Kolkata|           Low|              1|          68000.0|
|Bengaluru|          High|              1|         150000.0|
|   Mumbai|          High|              1|         140000.0|
|  Kolkata|        Medium|              1|          90000.0|
|  Gurgaon|          High|              1|         155000.0|
|Ahmedabad|        Medium|              1|         110000.0|
|  Chennai|           Low|              1|          72000.0|
|Bengaluru|        Medium|              1|          82000.0|
|    Delhi|          High|              1|         170000.0|
|  Unknown|        Medium|              2|          82500.0|
|    Delhi|        Medium|              1|         120000.0|
|    Delhi|           Low|              1|          65000.0|
|  Chennai|          High|              1|         160000.0|
+---------+--------------+---------------+-----------------+
```



```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("ELT") \
    .getOrCreate()
```

```
spark_df = spark.read.csv(
    "/content/drive/MyDrive/fde/lab3/customers.csv",          # EXTRACT
    header=True,
    inferSchema=True
)
```

```
spark_df.write.mode("overwrite").parquet("/content/raw_customers")        # LOAD
```

```
df = spark.read.parquet("/content/raw_customers")

from pyspark.sql.functions import col, upper, when, count, avg

df = df.dropDuplicates()
df = df.fillna({"city": "Unknown"})
df = df.withColumn("customer_name", upper(col("customer_name")))            # TRANSFORM

df = df.withColumn(
    "spend_category",
    when(col("annual_spend") < 80000, "Low")
    .when(col("annual_spend") <= 120000, "Medium")
    .otherwise("High")
)

elt_df = df.groupBy("city", "spend_category").agg(
    count("customer_id").alias("total_customers"),
    avg("annual_spend").alias("avg_spend")
)
elt_df.show()
```

```
+---------+--------------+---------------+-----------------+
|     city|spend_category|total_customers|        avg_spend|
+---------+--------------+---------------+-----------------+
|  Unknown|          High|              1|         132000.0|
|  Unknown|           Low|              3|71333.33333333333|
|     Pune|          High|              1|         145000.0|
|    Noida|           Low|              1|          78000.0|
|  Kolkata|           Low|              1|          68000.0|
|Bengaluru|          High|              1|         150000.0|
|   Mumbai|          High|              1|         140000.0|
|  Kolkata|        Medium|              1|          90000.0|
|  Gurgaon|          High|              1|         155000.0|
|Ahmedabad|        Medium|              1|         110000.0|
|  Chennai|           Low|              1|          72000.0|
|Bengaluru|        Medium|              1|          82000.0|
|    Delhi|          High|              1|         170000.0|
|  Unknown|        Medium|              2|          82500.0|
|    Delhi|        Medium|              1|         120000.0|
|    Delhi|           Low|              1|          65000.0|
|  Chennai|          High|              1|         160000.0|
+---------+--------------+---------------+-----------------+
```