Market Basket Insights

IBM Naan Mudhalvan Phase 3
Project Submission
Data Preprocessing

Table Of Contents

Loading The Dataset Into Jupyter Notebook

In the dataset preprocessing step, a parsing error occurred while loading the data into Jupyter Notebook, specifically with the 111th row. To resolve this issue, we identified and removed the problematic row, allowing for successful loading and analysis.

```
import pandas as pd
df=pd.read_csv("Assignment-1_Data.csv")
print(df)
```

```
---------------------------------------------------------------------------
ParserError                               Traceback (most recent call last)
Cell In[2], line 2
      1 import pandas as pd
----> 2 df=pd.read_csv("Assignment-1_data.csv")
      3 print(df)

File ~\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\_libs\parsers.pyx:859, in pandas._libs.parsers.TextReader._check_token
ize_status()

File ~\AppData\Local\Programs\Python\Python310\lib\site-packages\pandas\_libs\parsers.pyx:2025, in pandas._libs.parsers.raise_parser_error()

ParserError: Error tokenizing data. C error: Expected 2 fields in line 111, saw 3
```

[1]:
```
import pandas as pd
df=pd.read_csv("Assignment-1_Data.csv",delimiter=';')
pd.set_option('display.max_columns', None)
print(df)
```

```
        BillNo                            Itemname  Quantity  \
0       536365   WHITE HANGING HEART T-LIGHT HOLDER         6
1       536365                  WHITE METAL LANTERN         6
2       536365       CREAM CUPID HEARTS COAT HANGER         8
3       536365  KNITTED UNION FLAG HOT WATER BOTTLE         6
4       536365       RED WOOLLY HOTTIE WHITE HEART.         6
...        ...                                  ...       ...
522059  581587          PACK OF 20 SPACEBOY NAPKINS        12
522060  581587          CHILDREN'S APRON DOLLY GIRL         6
522061  581587         CHILDRENS CUTLERY DOLLY GIRL         4
522062  581587      CHILDRENS CUTLERY CIRCUS PARADE         4
522063  581587         BAKING SET 9 PIECE RETROSPOT         3

                  Date Price  CustomerID         Country
0       01.12.2010 08:26  2,55     17850.0  United Kingdom
1       01.12.2010 08:26  3,39     17850.0  United Kingdom
2       01.12.2010 08:26  2,75     17850.0  United Kingdom
3       01.12.2010 08:26  3,39     17850.0  United Kingdom
4       01.12.2010 08:26  3,39     17850.0  United Kingdom
...                  ...   ...         ...             ...
522059  09.12.2011 12:50  0,85     12680.0          France
522060  09.12.2011 12:50   2,1     12680.0          France
522061  09.12.2011 12:50  4,15     12680.0          France
522062  09.12.2011 12:50  4,15     12680.0          France
522063  09.12.2011 12:50  4,95     12680.0          France

[522064 rows x 7 columns]
```

After removing the problematic 111th row, the dataset was successfully loaded into Jupyter Notebook without any parsing errors. The dataset is now viewable after using the pandas **read_csv** function and ready for analysis. The dataset consists of **522064 rows x 7 columns**

Data Cleaning: Fixing Unnecessary Characters in Entries

```
[3]: column_name = 'Country,,,'
     print(df[column_name])

0           United Kingdom,,
1           United Kingdom,,
2           United Kingdom,,
3           United Kingdom,,
4           United Kingdom,,
                  ...
522058             France,,
522059             France,,
522060             France,,
522061             France,,
522062             France,,
Name: Country,,,, Length: 522063, dtype: object

[4]: df['Country,,,'] = df['Country,,,'].str.replace(',', '')

[5]: df.head()
```

[5]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country,,,
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2,55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3,39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2,75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3,39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3,39	17850.0	United Kingdom

We identified and corrected discrepancies in the **'Country'** column, where entries contained unnecessary **',,,'** characters. By eliminating these inconsistencies, we established uniform data.

```
[8]: df['Price'] = df['Price'].str.replace(',', '.').astype(float)
```

```
[9]: df.head()
```

[9]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3.39	17850.0	United Kingdom

Discovered irregular numerical representation in the **'Price'** column, where numbers used commas instead of decimal points . We corrected this format inconsistency, converting all entries into the appropriate float format

Removing Rows Containing Null Values

```
[10]: missing_values = df.isnull().sum()
      print(missing_values)
```

```
BillNo          0
Itemname     2154
Quantity        2
Date         1231
Price        1231
CustomerID  134573
Country      1231
dtype: int64
```

```
[12]: df1 = df.dropna()
```

```
[14]: df1
```

[14]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
522058	581587	PACK OF 20 SPACEBOY NAPKINS	12	09.12.2011 12:50	0.85	12680.0	France
522059	581587	CHILDREN'S APRON DOLLY GIRL	6	09.12.2011 12:50	2.10	12680.0	France
522060	581587	CHILDRENS CUTLERY DOLLY GIRL	4	09.12.2011 12:50	4.15	12680.0	France
522061	581587	CHILDRENS CUTLERY CIRCUS PARADE	4	09.12.2011 12:50	4.15	12680.0	France
522062	581587	BAKING SET 9 PIECE RETROSPOT	3	09.12.2011 12:50	4.95	12680.0	France

387490 rows × 7 columns

Using the **'isnull'** function, we systematically identified and counted null values across the dataset. Following this, we applied the **'dropna'** function to remove rows containing any null values within a single column. This is necessary in order to apply association rules efficiently. After this step the dataset contained only **387490 rows x 7 columns.**

Preparing Dataset for Applying Association Rules

```
transaction_data = df1.groupby(['BillNo', 'Date'])['Itemname'].agg(lambda x: ','.join(x)).reset_index()
```

```
transaction_data
```

	BillNo	Date	Itemname
0	536365	01.12.2010 08:26	WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL...
1	536366	01.12.2010 08:28	HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT
2	536367	01.12.2010 08:34	ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUS...
3	536368	01.12.2010 08:34	JAM MAKING SET WITH JARS,RED COAT RACK PARIS F...
4	536369	01.12.2010 08:35	BATH BUILDING BLOCK WORD
...	...	...	...
18184	581583	09.12.2011 12:23	LUNCH BAG RED RETROSPOT,6 CHOCOLATE LOVE HEART...
18185	581584	09.12.2011 12:25	RED FLOCK LOVE HEART PHOTO FRAME,6 CHOCOLATE L...
18186	581585	09.12.2011 12:31	BLACK TEA TOWEL CLASSIC DESIGN,ASSORTED BOTTLE...
18187	581586	09.12.2011 12:49	LARGE CAKE STAND HANGING STRAWBERY,SET OF 3 H...
18188	581587	09.12.2011 12:50	CIRCUS PARADE LUNCH BOX,PLASTERS IN TIN CIRCUS...

18189 rows × 3 columns

```
[17]: transaction_data = transaction_data.drop(['BillNo', 'Date'], axis=1)
```

```
[18]: transaction_data
```

[18]:

	Itemname
0	WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL...
1	HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT
2	ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUS...
3	JAM MAKING SET WITH JARS,RED COAT RACK PARIS F...
4	BATH BUILDING BLOCK WORD
...	...
18184	LUNCH BAG RED RETROSPOT,6 CHOCOLATE LOVE HEART...
18185	RED FLOCK LOVE HEART PHOTO FRAME,6 CHOCOLATE L...
18186	BLACK TEA TOWEL CLASSIC DESIGN,ASSORTED BOTTLE...
18187	LARGE CAKE STAND HANGING STRAWBERY,SET OF 3 H...
18188	CIRCUS PARADE LUNCH BOX,PLASTERS IN TIN CIRCUS...

18189 rows × 1 columns

Combined items with the same bill number and date, optimizing the dataset for applying association rules. Dropped '**BillNo**' and '**Date**' columns as they are unnecessary for association rule analysis, ensuring a streamlined dataset. After performing this step the dimensions of the dataset were **18189 rows x 3 columns**. Now the dataset is ready for applying association rules

Summary Of All Preprocessing Steps

Cleaned 'Country' Names:
Removed trailing '„' characters, ensuring all country names were standardized.

Formatted 'Price' Data:
Transformed irregular 'Price' formats into floats for precise calculations.

Ensured Data Completeness:
Eliminated missing values with '**isnull**' and '**dropna**' methods, resulting in a complete dataset.

Optimized for Analysis:
Grouped items with the same bill number and date, simplifying the dataset for efficient analysis.

Enhanced Focus:
Trimmed unnecessary 'Bill Number' and 'Date' columns to create a streamlined dataset, ready for in-depth analysis.