

Market Basket Insights

IBM Naan Mudhalvan Phase 3
Project Submission
Data Preprocessing

Table Of Contents

O1 Loading The Dataset Into Jupyter Notebook

Data Cleaning: Fixing Unnecessary Characters in Entries

03 Removing Rows Containing Null Values

O4 Preparing Dataset for Applying Association Rules



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.

```
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
                                                          Ouantity \
                                                Itemname
             536365
                      WHITE HANGING HEART T-LIGHT HOLDER
             536365
                                     WHITE METAL LANTERN
                          CREAM CUPID HEARTS COAT HANGER
             536365
             536365 KNITTED UNION FLAG HOT WATER BOTTLE
             536365
                          RED WOOLLY HOTTIE WHITE HEART.
                                                               . . .
                             PACK OF 20 SPACEBOY NAPKINS
     522059 581587
                                                                12
     522060
             581587
                             CHILDREN'S APRON DOLLY GIRL
     522061
                            CHILDRENS CUTLERY DOLLY GIRL
     522062 581587
                         CHILDRENS CUTLERY CIRCUS PARADE
     522063 581587
                            BAKING SET 9 PIECE RETROSPOT
                         Date Price CustomerID
                                                        Country
             01.12.2010 08:26 2.55
                                        17850.0 United Kingdom
             01.12.2010 08:26 3,39
                                        17850.0 United Kingdom
             01.12.2010 08:26 2,75
                                        17850.0 United Kingdom
             01.12.2010 08:26 3.39
                                        17850.0 United Kingdom
             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
                                        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**

<u>Data Cleaning: Fixing Unnecessary Characters</u> <u>in Entries</u>

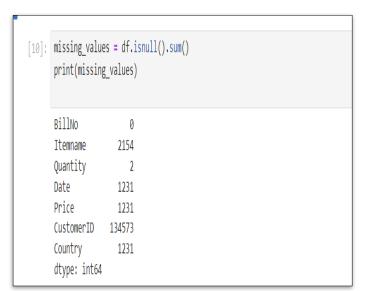
pr	rint(df[corami_name)						
0		United King							
1		United King							
2		United King							
3		United King							
4									
	22058								
	22050		ince,,						
	22060		ince,,						
	22061		ince,,						
	**								
	22062	Fra	ince						
52 Na	ame: Cou	ntry,,,, Ler	ngth: 52206	53, dtype: ob	_	, '')			
52 Na df	ame: Cou	ntry,,,, Ler ry,,,'] = df	ngth: 52206		_	, '')			
52 Na df	ame: Cou f['Count	ntry,,,, Ler ry,,,'] = df	ngth: 52206		olace(','		Price	CustomerID	Country,,,
52 Na df	ame: Count f['Count f.head() BillNo	ntry,,,, Ler	gth: 52206	,,,'].str.rep	Quantity		Price 2,55	CustomerID 17850.0	
52 Na df df	ame: Cou f['Count f.head() BillNo	ntry,,,, Ler	ngth: 52206 ['Country,	.,,'].str.rep	Quantity	Date			United Kingdom
52 Na df df	ame: Count f['Count f.head() BillNo 536365	ntry,,,, Lerry,,,'] = df	ngth: 52206 ['Country, NG HEART T- WHITE M	Itemname	Quantity	Date 01.12.2010 08:26 01.12.2010 08:26	2,55	17850.0	United Kingdom United Kingdom
52 Na df df 0 1 2	f ('Count f ('Count f head() BillNo 536365 536365	white HANG	ngth: 52206 ['Country, ING HEART T- WHITE M UPID HEARTS	Itemname LIGHT HOLDER	Quantity 6 6 8	Date 01.12.2010 08:26 01.12.2010 08:26	2,55 3,39	17850.0 17850.0 17850.0	United Kingdom United Kingdom

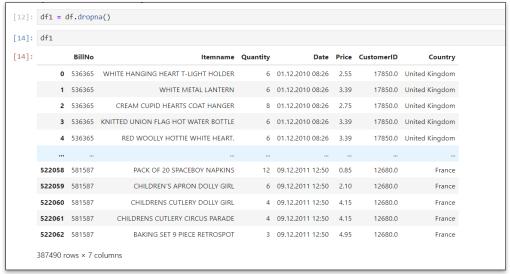
We identified and corrected discrepancies in the **'Country'** column, where entries contained unnecessary **',,,'** characters. By eliminating these inconsistencies, we established uniform data.

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

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

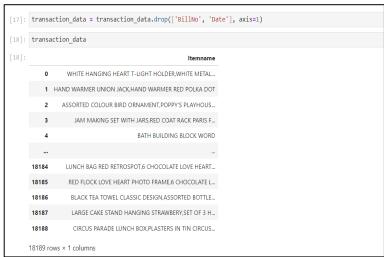




Using the **'isnull'** function, we systematically identified and counted null values across the dataset. Following this, we applied the **'dropna'** function to remove rows containing any null values within a single column. This is necessary in order to apply association rules efficiently. After this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

<u>Preparing Dataset for Applying Association</u> Rules

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				



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

<u>Summary Of All Preprocessing Steps</u>



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.