Product Recomendation System (Basket Analysis)

Presented By:

- Alidzhon Aminov
- Hernan Mauricio Leon Barreto
- Omar Hany Sheiba Ibrahim Badr

OVERVIEW

- Problem Statement
- Type of method used
- Why basket Analysis ?
- Introduce Data Set
- Work Flow of the project
- Issues in the data set
- Interpretation of Results
- Conclusions

Problem Statement

In the competitive e-commerce sector, businesses face the challenge of enhancing customer experiences and increasing sales by effectively predicting and understanding consumer purchasing behaviors. Market Basket Analysis (MBA) addresses this by uncovering product associations, facilitating smarter cross-selling strategies, optimizing inventory management, and enabling more targeted marketing efforts

Data Information

Source: UCI Machine Learning Repository by School of Engineering, London South Bank University.

Content: Transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail

Type of method used

Market Basket Analysis (MBA) utilized to discover relationships between items within extensive datasets, making it highly valuable in retail environments for elucidating consumer purchase patterns. In addressing the problem of enhancing customer experiences and increasing sales in e-commerce, this MBA have the next strategic objectives:

- Discovery of Product Associations
- Optimization of Cross-Selling Strategies
- Enhanced Inventory Management
- Refinement of Marketing Campaigns

Why basket Analysis ?

Basket analysis is chosen in e-commerce because it offers:

- Deep insights into customer behavior.
- Effective and targeted marketing strategies.
- Improved product placement and store layout.
- Enhanced customer segmentation.

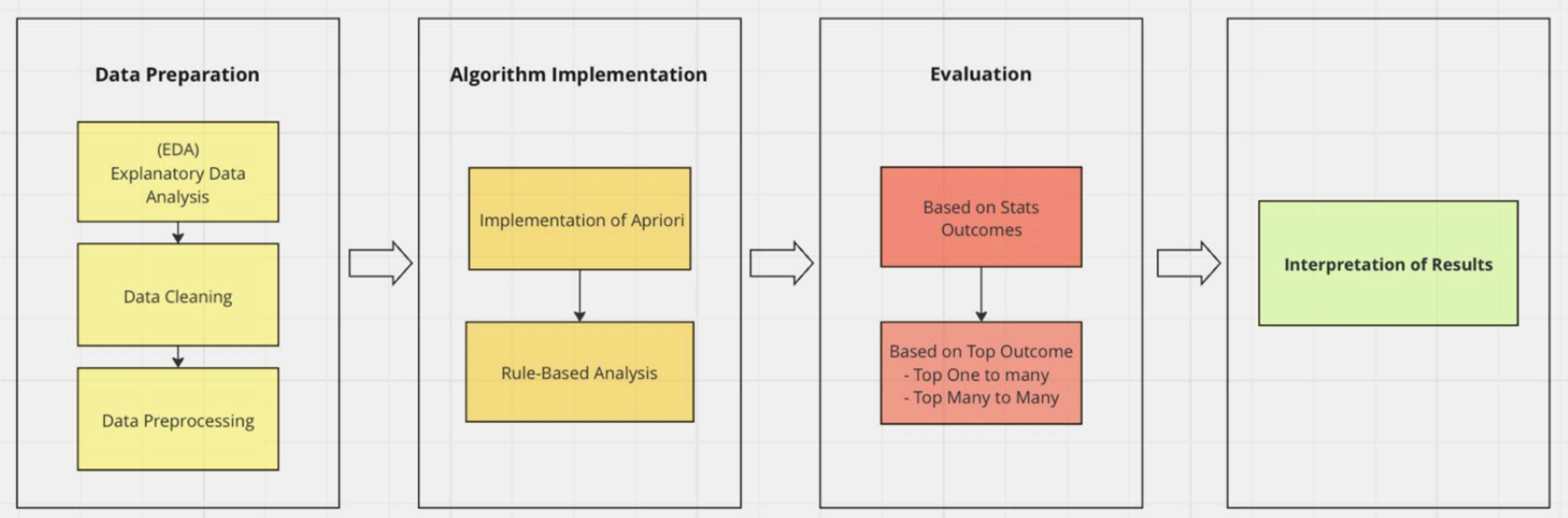
Introduce Dataset

Features

- InvoiceNo: A 6-digit number uniquely assigned to each transaction. If the number is prefixed with 'c', it indicates a cancellation. (Nominal)
- StockCode: A unique identifier for each product sold by the retailer. (Nominal)
- Description: The name or a brief description of the product. (Nominal)
- Quantity: The number of units of the product sold in each transaction. (Numeric)
- InvoiceDate: The date and time when the transaction was made. (Datetime)
- UnitPrice: The price per unit of the product in sterling. (Numeric)
- Country The country where the customer resides. (Nominal)

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	12/1/2010 9:41	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	12/1/2010 9:49	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	12/1/2010 10:24	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	12/1/2010 10:24	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	12/1/2010 10:24	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	12/9/2011 9:57	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	12/9/2011 10:28	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	12/9/2011 11:57	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	12/9/2011 11:58	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	12/9/2011 11:58	1.25	17315.0	United Kingdom

Work flow of the project



Issues in the data set and the process

Exploratory Data Analysis (EDA) - Data Understanding and Preparation

- Negative Prices: Presence of negative prices likely due to special cases (refund, damage...).
- Missing Values: Missing values in critical columns like Description.
- Inconsistent in the product descriptions.
- Right-Skewed Distribution of Items per Invoice: Highly right-skewed distribution of the number of items per invoice.

Data Preprocessing for Machine Learning

- Handling Rare Items: Presence of rarely purchased items complicating the analysis.
- Class Imbalance: Imbalanced class distribution.
- Scaling and Normalization: Features with different scales affecting the model disproportionately.

Association Rule Mining

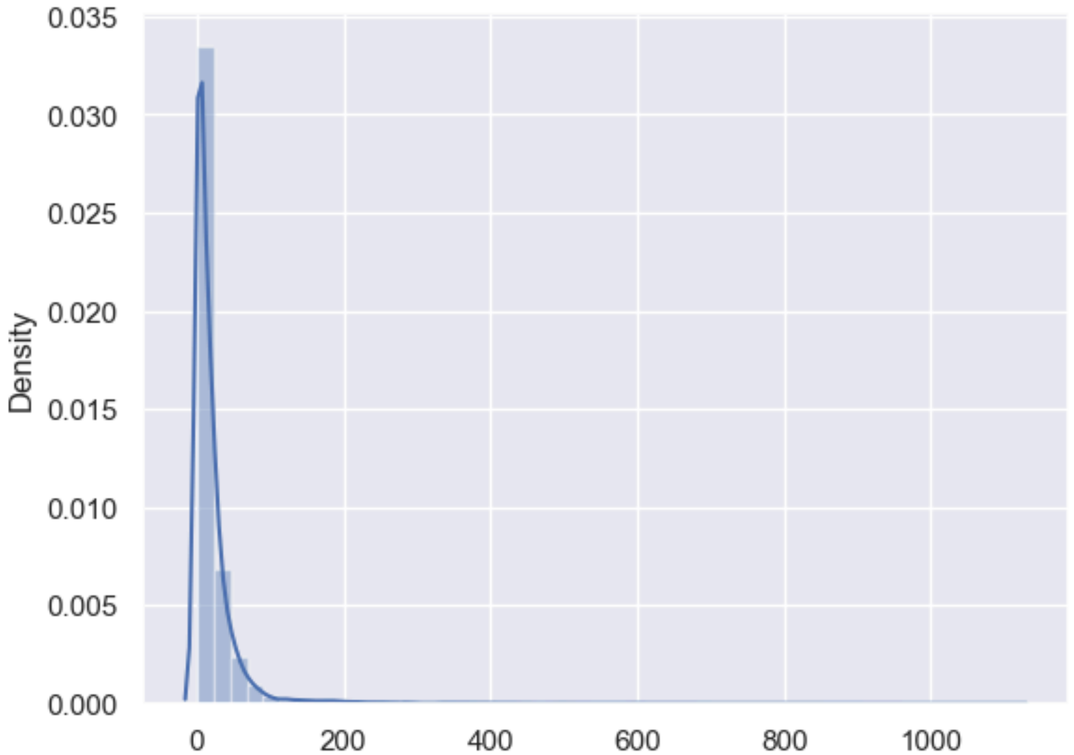
- Threshold Selection for Support, Confidence, and Lift: Determining the right thresholds for meaningful association rules.

Feature Engineering and Encoding

- Encoding Categorical Variables: Need to convert categorical variables to numerical values.

Model Training and Evaluation

- Hyperparameter Tuning Finding the optimal hyperparameters for the machine learning model.

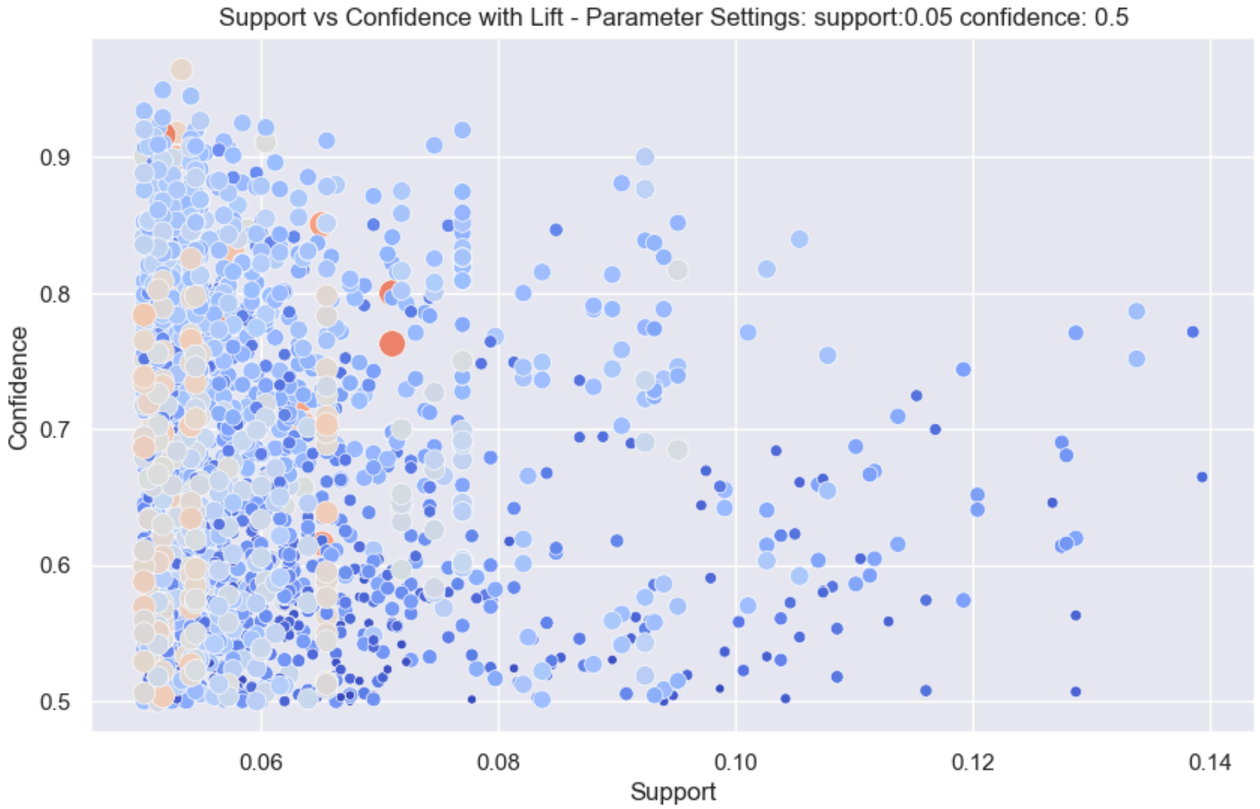



```
data = list(df[(df['Description'].str.len() < 20) & (df['Description'].isin(filtered_uppercase_descriptions) != True)]['Descrip
data.append("DOTCOM POSTAGE")
```

```
onsistent_data, return_counts=True)
```

```
'?', '? sold as sets?', '??', '?? missing', '???',
ges????', '????missing', '???lost', '???missing',
?', '?lost', '?missing', '?sold as sets?',
ad debt', 'Adjustment', 'Amazon', 'Amazon Adjustment',
old sets', 'Bank Charges', 'Breakages', 'CRUK Commission',
, 'DOTCOM POSTAGE', 'Dagamed', 'Damaged', 'Damages',
samples', 'Discount', 'Display', 'Dotcom', 'Dotcom sales',
et', "Dotcom sold in 6's", 'Found', 'Found by jackie',
 w/hse', 'Given away', 'Had been put aside.',
is', 'Manual', 'Marked as 23343', 'Missing',
 Carriage', 'OOPS ! adjustment', 'Sale error',
ples', 'Sold as 1 on dotcom', 'Thrown away-rusty',
way.', 'Unknown', 'Water damaged', 'adjust', 'adjustment',
 'amazon adjust', 'amazon sales', 'barcode problem',
 'came coded as 20713', "can't find", 'check', 'check?',
 up? 84930', 'counted', 'cracked', 'crushed',
boxes', 'crushed ctn', 'damaged', 'damaged stock',
, 'damages wax', 'damages/display', 'damages/dotcom?',
', 'dotcom', 'dotcom adjust', 'dotcom sales',
old sets', 'dotcomstock', 'ebay', 'faulty', 'found',
x', 'had been put aside', 'label mix up', 'lost',
space', 'lost??', 'mailout', 'michel oops', 'missing',
', 'mix up with c', 'mixed up', 'mouldy',
unsaleable.', 'on cargo order', 're-adjustment',
', 'rusty throw away', 'rusty thrown away', 'samples',
```

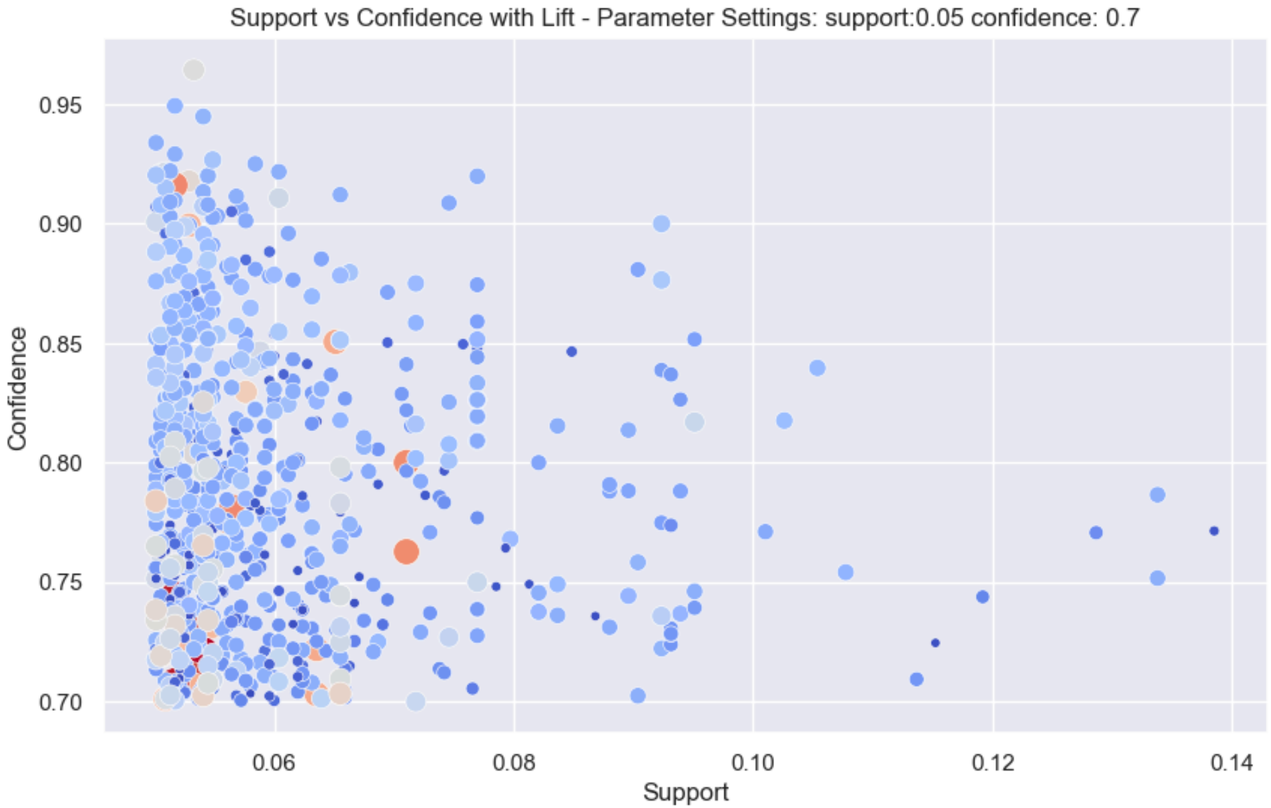
Interpretation of Results



Graph with Support: 0.05, Confidence: 0.5

According to the graph, the low support and moderate confidence helped to identify a large number of association rules. The lift (used as the hue in the graph) shows there are a lot of weak and moderate associations, and some associations are strong.

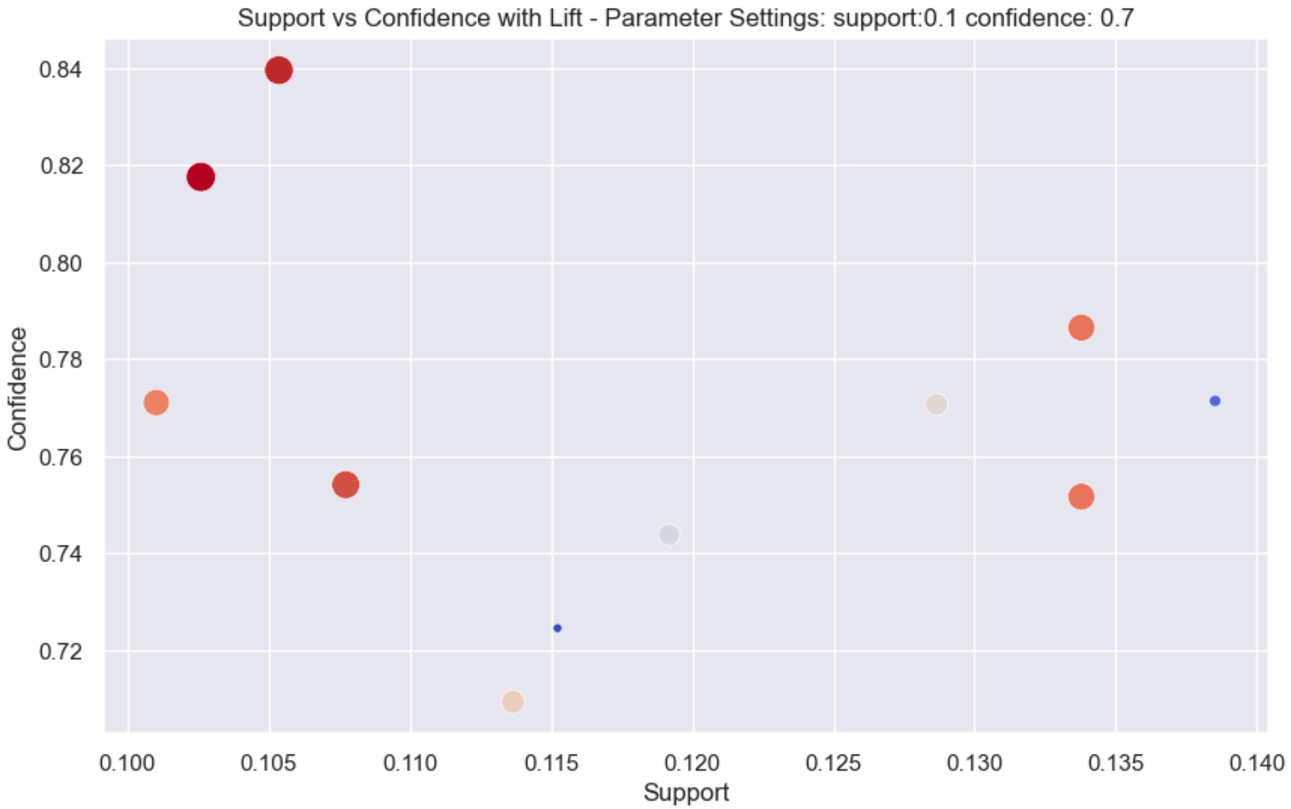
Interpretation of Results



Graph with Support: 0.05, Confidence: 0.7

It shows that increasing the confidence filters out some rules, but in general, the number of rules remains high due to the low support threshold. By increasing the confidence, we got rid of the rules that have reliability lower than 0.7.

Interpretation of Results



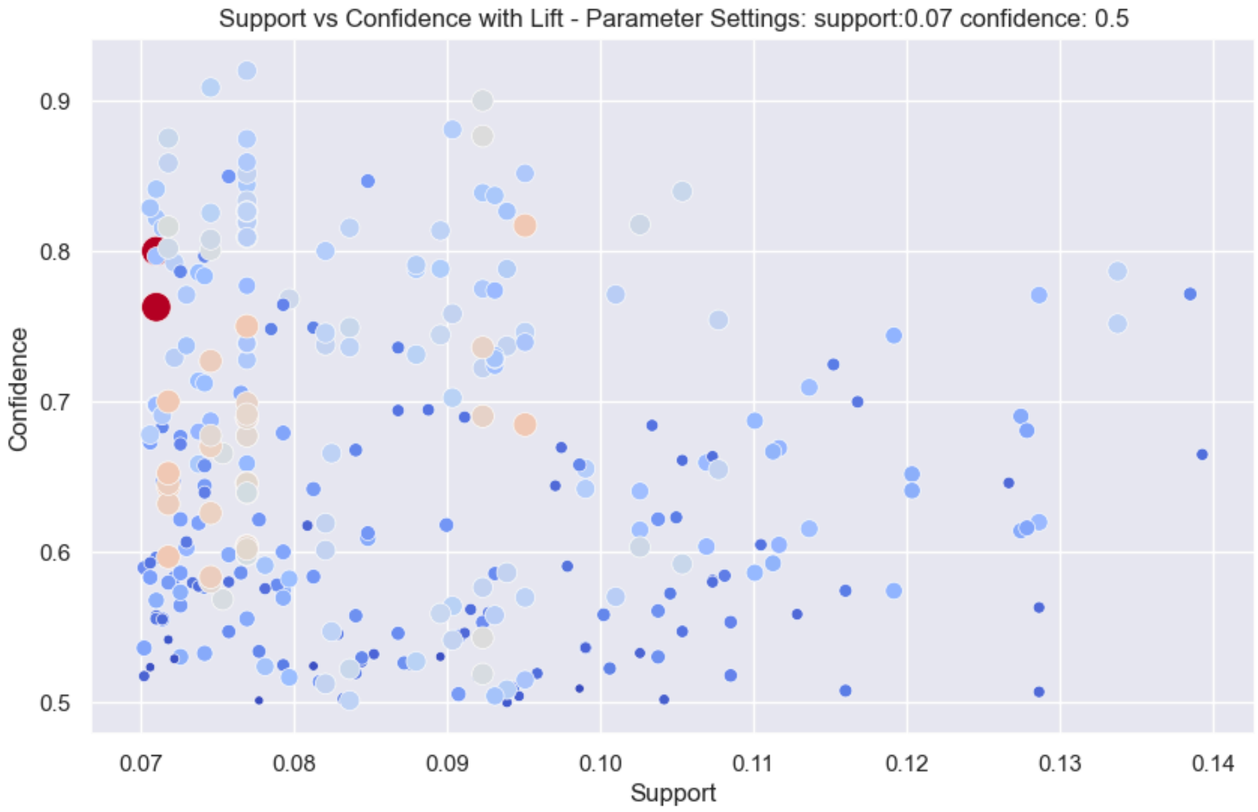
Graph with Support: 0.1, Confidence: 0.7

In this case, increasing the support while keeping the confidence level at 0.7, we have created a more strict rule which filtered out most of the rules, keeping only the top frequent ones.

For this dataset, such a setting gets rid of most of the association rules, which is not useful.

Also, increasing the confidence level does not have a dramatic effect on the top list as it only filters out the rules on the bottom. This means setting the confidence level that interests us and playing with support can help uncover a new set of rules which might help to uncover hidden patterns of association rules.

Interpretation of Results



The last graph with Support: 0.06, Confidence: 0.7 and the Top Singular and Top Multi proves that increasing the support helps to get more frequent association rules and uncover other rules patterns.

Conclusions

- Support Threshold: Higher support filters out less frequent itemsets, resulting in fewer but more meaningful rules, focusing on stronger patterns.

- Confidence Threshold: Higher confidence improves rule reliability but doesn't drastically change the top rules if they already have high confidence. Support plays a more critical role in shaping the overall rule set.

- The support and confidence thresholds should be set based on the specific needs and goals of the users.

Example – Associations

```
cart_items = {'CHARLOTTE BAG PINK POLKADOT',
              'CHARLOTTE BAG SUKI DESIGN',
              'PACK OF 72 RETROSPOT CAKE CASES',
              'STRAWBERRY CHARLOTTE BAG'}
recommend_items(cart_items, rules_005_05)
```

	item	support	confidence	lift	conviction
0	RED RETROSPOT CHARLOTTE BAG	0.119179	0.743842	3.583454	3.093498
1	WOODLAND CHARLOTTE BAG	0.113654	0.709360	3.840849	2.805225
2	CHARLOTTE BAG SUKI DESIGN	0.110103	0.687192	3.658287	2.596337
3	CHARLOTTE BAG PINK POLKADOT	0.102605	0.640394	3.836309	2.316620
4	STRAWBERRY CHARLOTTE BAG	0.102605	0.614657	3.836309	2.179304
5	BOX OF 24 COCKTAIL PARASOLS	0.052881	0.558333	4.112839	1.956784
6	REGENCY CAKESTAND 3 TIER	0.055643	0.542308	2.223637	1.652020
7	LUNCH BAG CARS BLUE	0.055643	0.542308	2.815999	1.764109
8	LUNCH BAG RED RETROSPOT	0.085241	0.532020	2.096638	1.594621
9	PACK OF 72 RETROSPOT CAKE CASES	0.085241	0.532020	2.146716	1.607270

Charlotte Bag Pink Polkadot

Charlotte Bag suki design

Retrospot cake cases

Strawberry Charlotte Bag

Woodland Charlotte Bag

Box of 24 COCTAIL PARASOL

Regency Cakestand 3 tier

Lunch Bag Cars Blue

RED RETROSPOT CHARLOTTE BAG

Lunch Bag Red Retrospot

Example – Associations

```
recommend_items(cart_items, rules_005_05)
```

[92]:

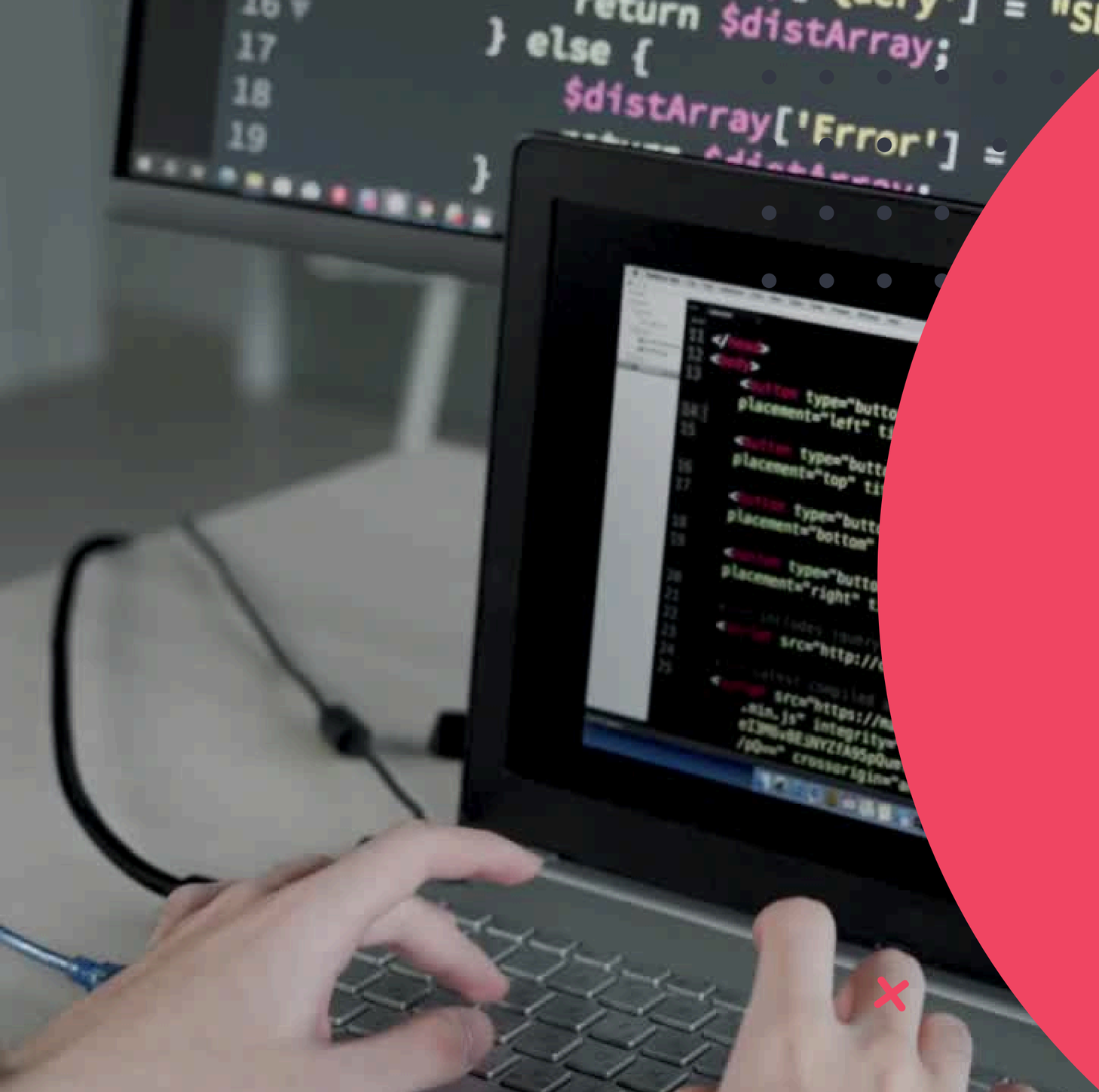
	item	support	confidence	lift	conviction
0	SMALL DOLLY MIX DESIGN ORANGE BOWL	0.058800	0.846591	6.094493	5.613026
1	SMALL MARSHMALLOWS PINK BOWL	0.057616	0.829545	7.125655	5.183689

[93]:

```
cart_items = {'JUMBO BAG SCANDINAVIAN BLUE PAISLEY'}
recommend_items(cart_items, rules_005_05)
```

[93]:

	item	support	confidence	lift	conviction
0	JUMBO BAG RED RETROSPOT	0.092344	0.774834	2.671334	3.152990
1	JUMBO SHOPPER VINTAGE RED PAISLEY	0.073796	0.619205	3.157075	2.111026
2	JUMBO STORAGE BAG SUKI	0.070245	0.589404	2.812711	1.925128
3	JUMBO BAG PINK VINTAGE PAISLEY	0.069850	0.586093	3.676136	2.030813
4	JUMBO BAG PINK POLKADOT	0.067088	0.562914	3.134997	1.877072
5	JUMBO BAG APPLES	0.064325	0.539735	3.335826	1.821126
6	JUMBO BAG BAROQUE BLACK WHITE	0.061168	0.513245	3.227203	1.727692



Run the code