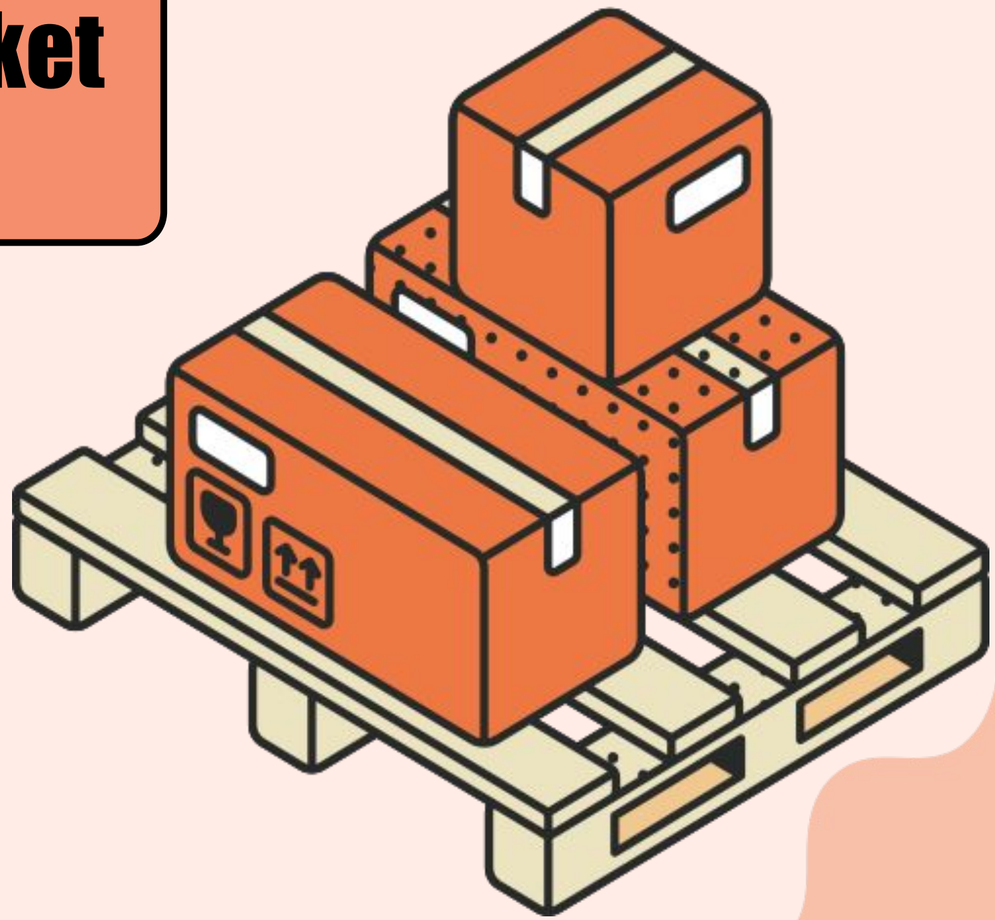
Uncovering Cross-Category Purchase Patterns using Market Basket Analysis



Marketing Analytics

Jagruta Advani (ja53837)
Navya Singhal (ns38323)
Sarthak Shivnani (ss223347)
Sushanth Ravichandran (sr’56925)
Vishwa Patel (vp8792)

Agenda

01 Introduction

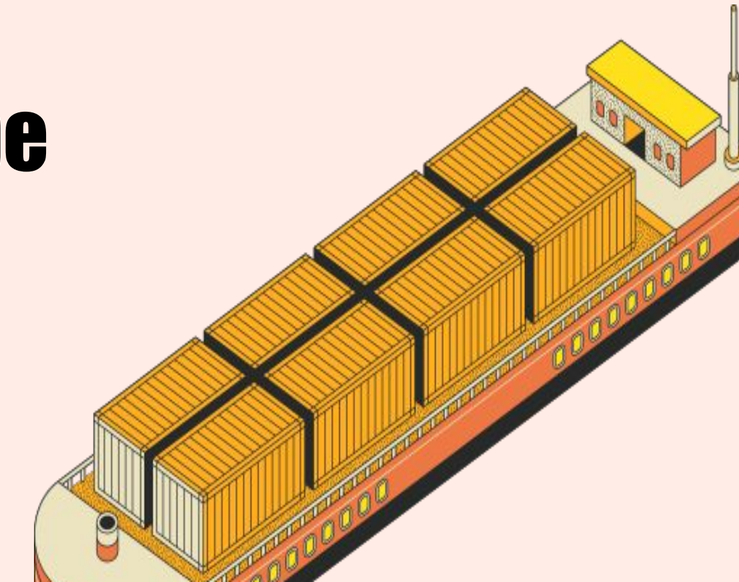
02 Exploratory Analysis

03 Market Basket Analysis

04 (dis)Honorable Mentions

05 Results

06 Future Scope



The Dataset

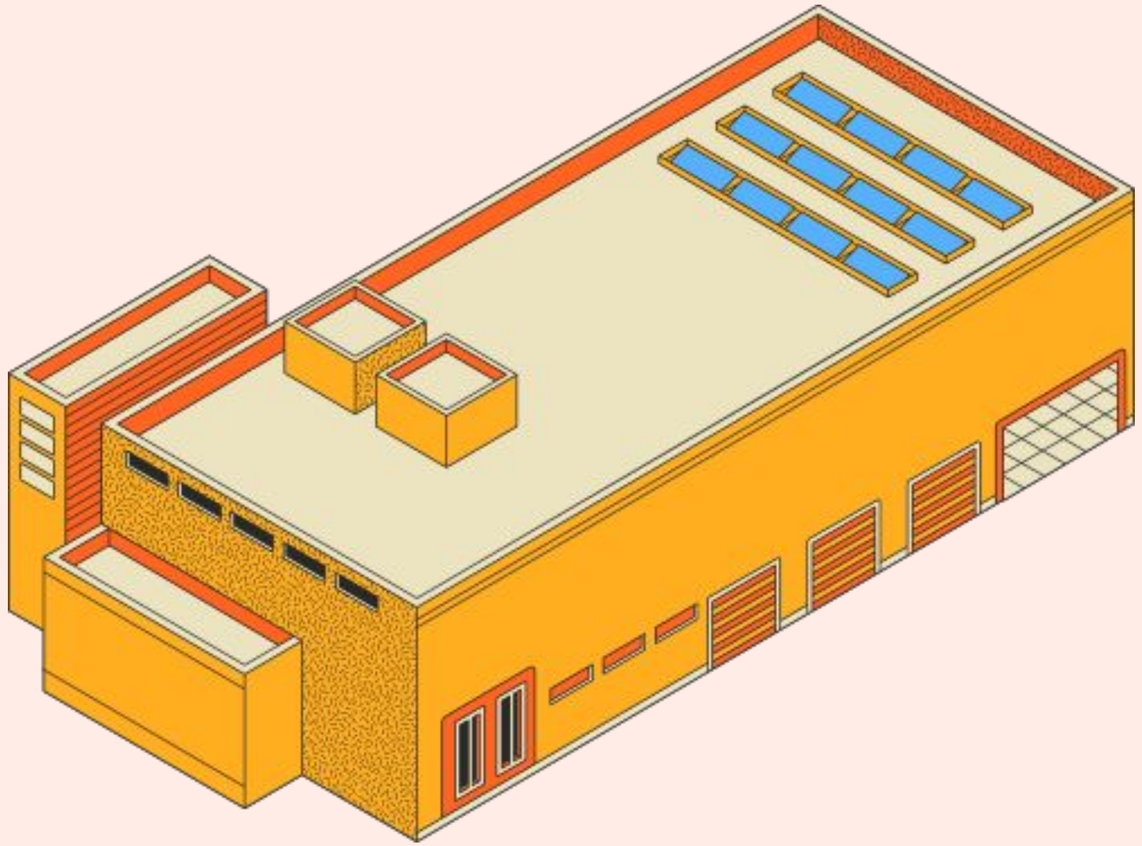
UK-based online retailer transaction data mainly selling **unique all-occasion gifts**
Transaction data ranges from **01/12/2010 and 09/12/2011**

Includes **541,909 transactions** with **6 features**
Contains multivariate, sequential, and time-series data

Variable	Role	Type	Description
InvoiceNo	ID	Categorical	Number uniquely assigned to each transaction
StockCode	ID	Categorical	Number uniquely assigned to each distinct product
Description	Feature	Categorical	Product name
Quantity	Feature	Integer	Quantities of each product (item) per transaction
InvoiceDate	Feature	Date	Day and time when each transaction was generated
UnitPrice	Feature	Continuous	Product price per unit
CustomerID	Feature	Categorical	Number uniquely assigned to each customer
Country	Feature	Categorical	Name of the country where each customer resides

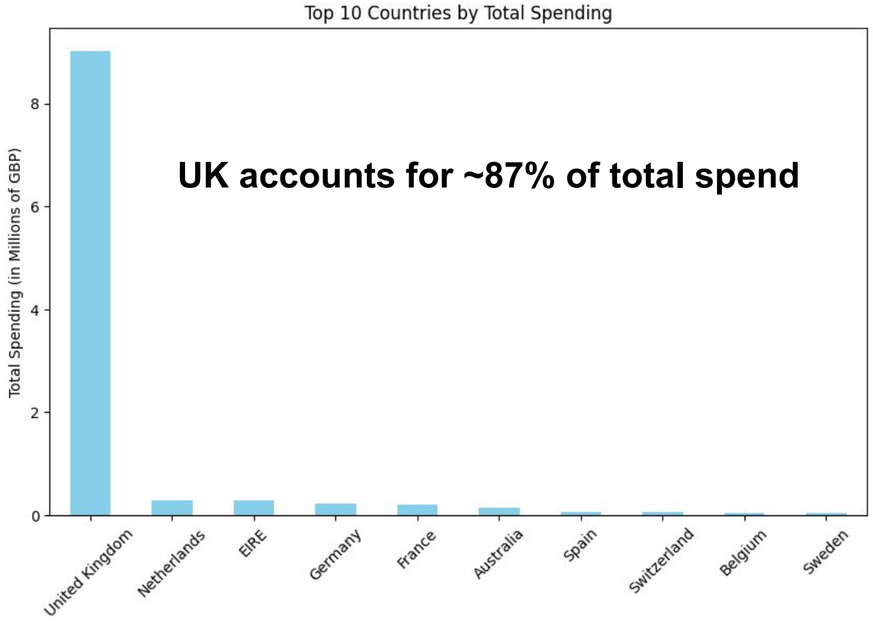
Source: UC Irvine Machine Learning Repository

“How can we identify product bundles that maximize customer purchase and average order value (AOV)?”

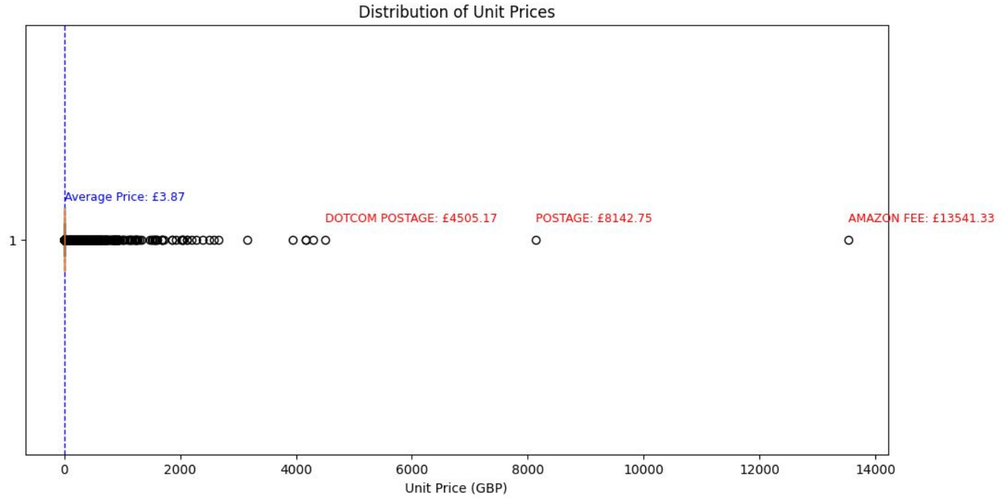


Exploratory Analysis

Overview



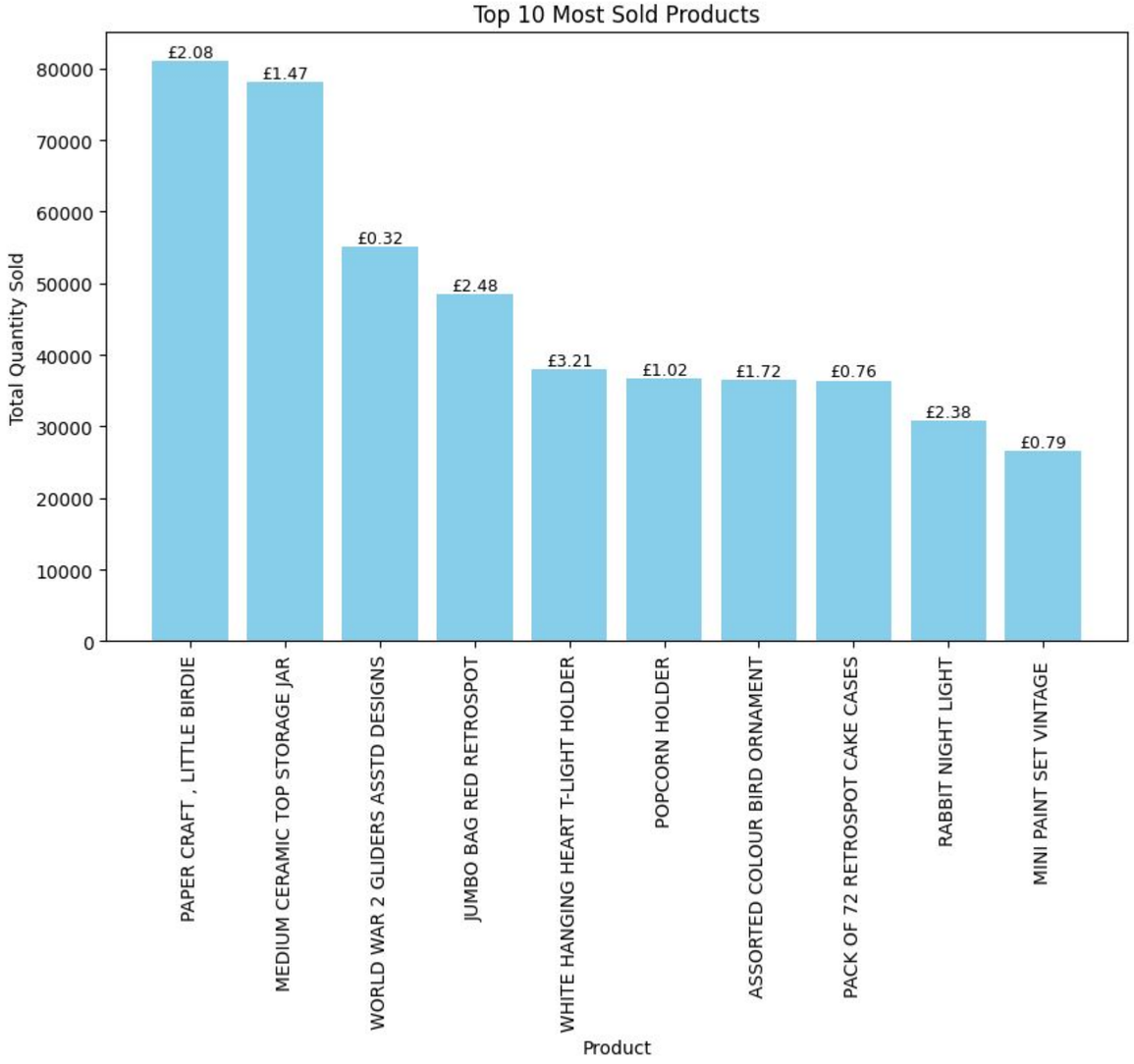
UK accounts for ~87% of total spend



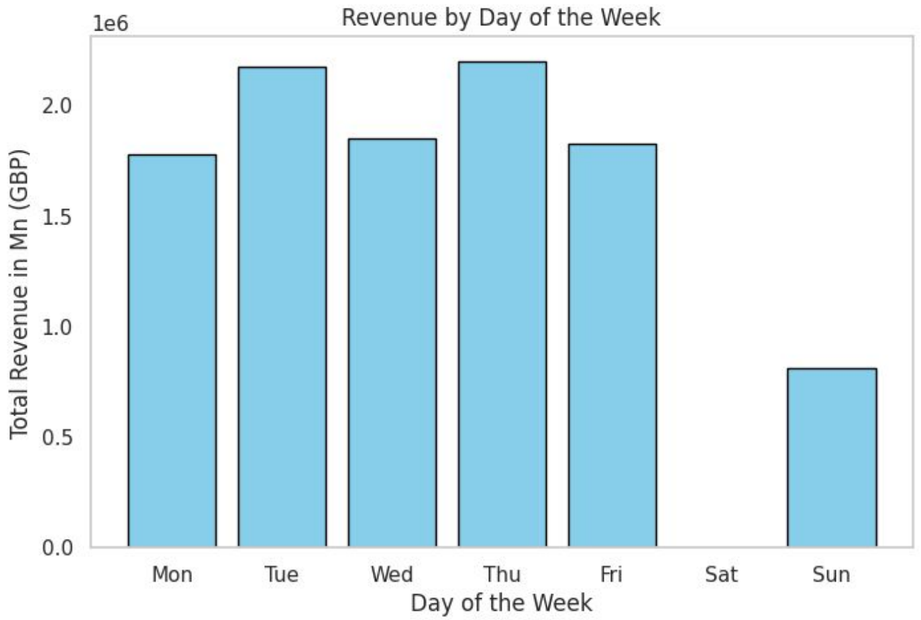
Average price of products sold is £3.9

(PICNIC BASKET WICKER 60pc as the highest priced product at £650)

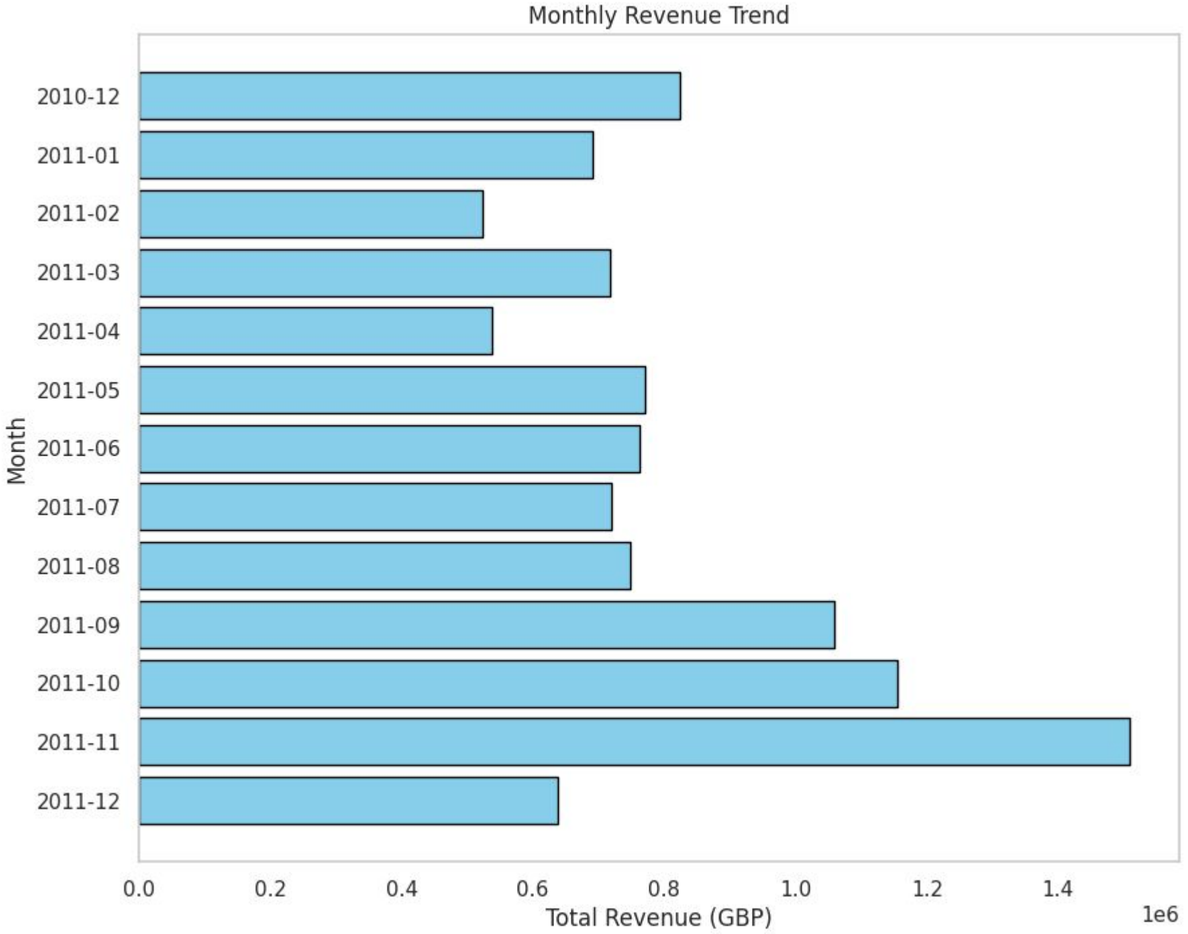
High volume sales driven by affordable, frequently purchased goods



Impact of Time on Sales



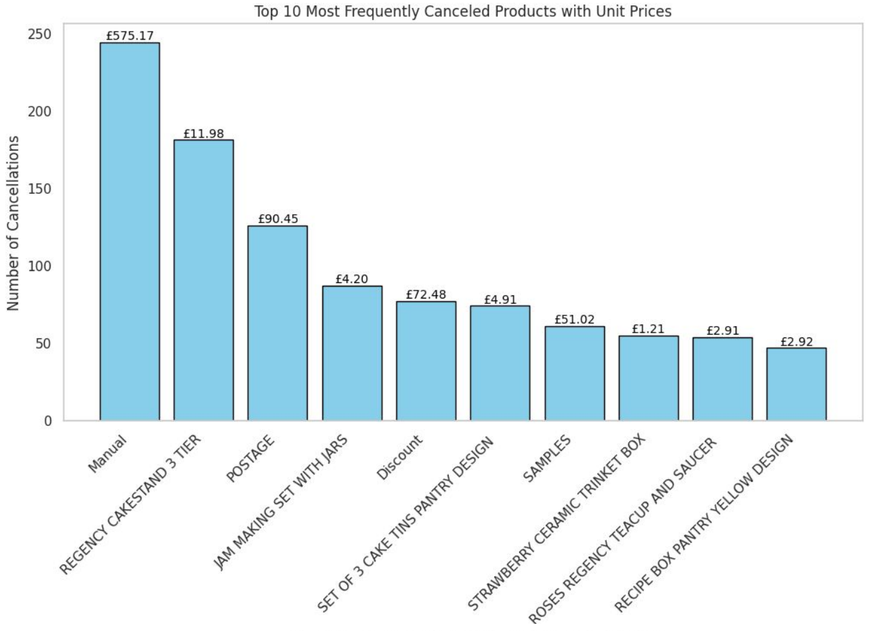
No sales on Saturday might indicate lack of purchase orders placed on weekends from commercial customers

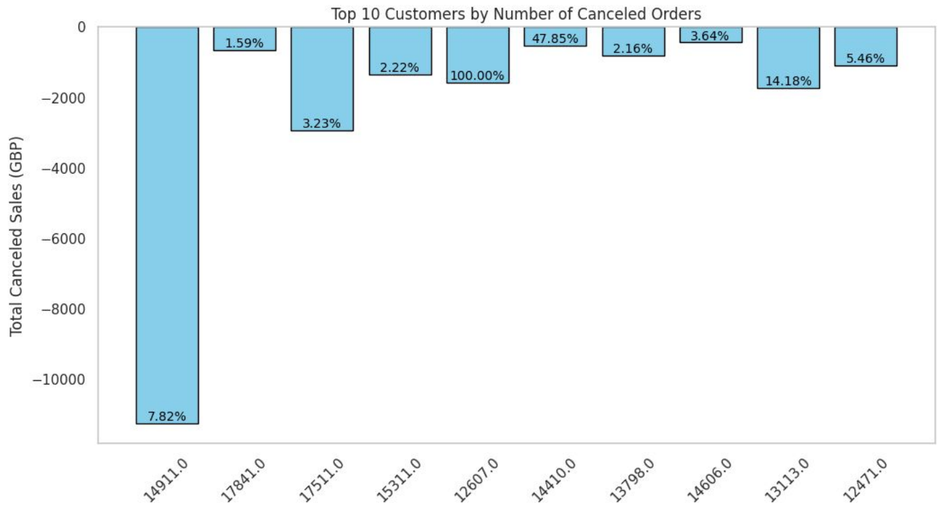


Peak sales during the holiday season

(Note: Dec 2011 has data for 9 days only)

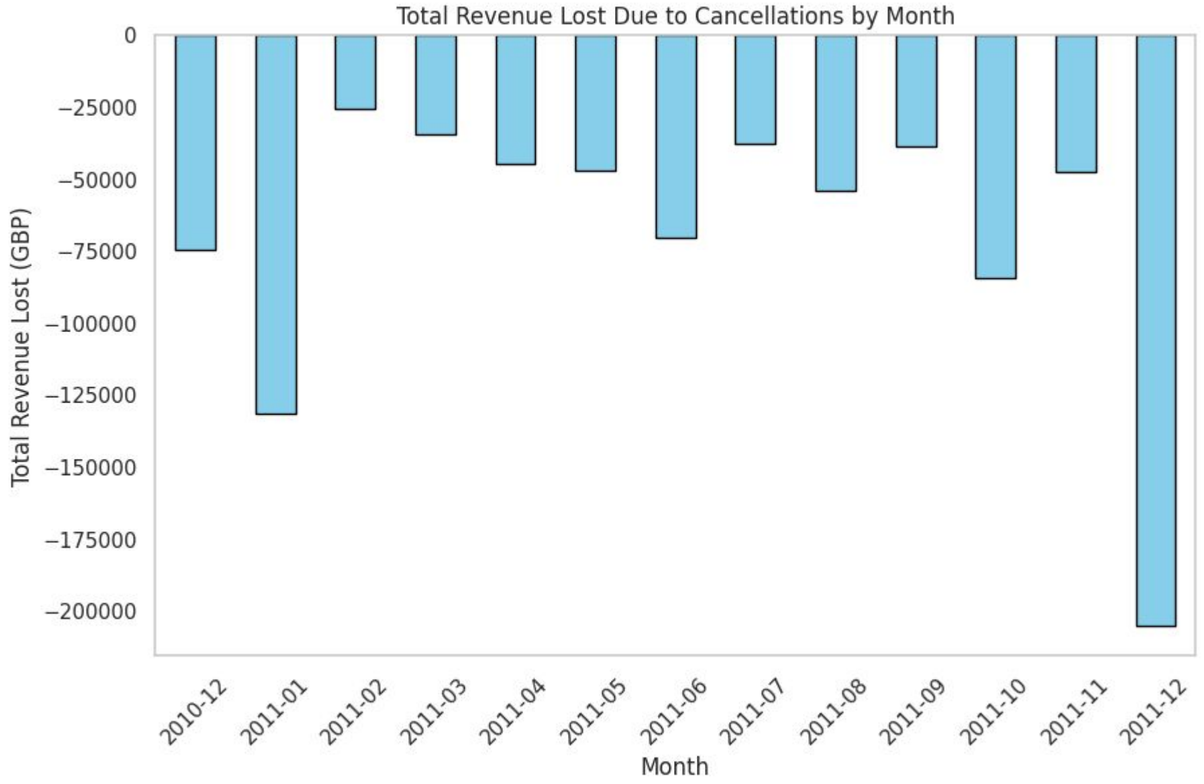
Trend of Cancellations





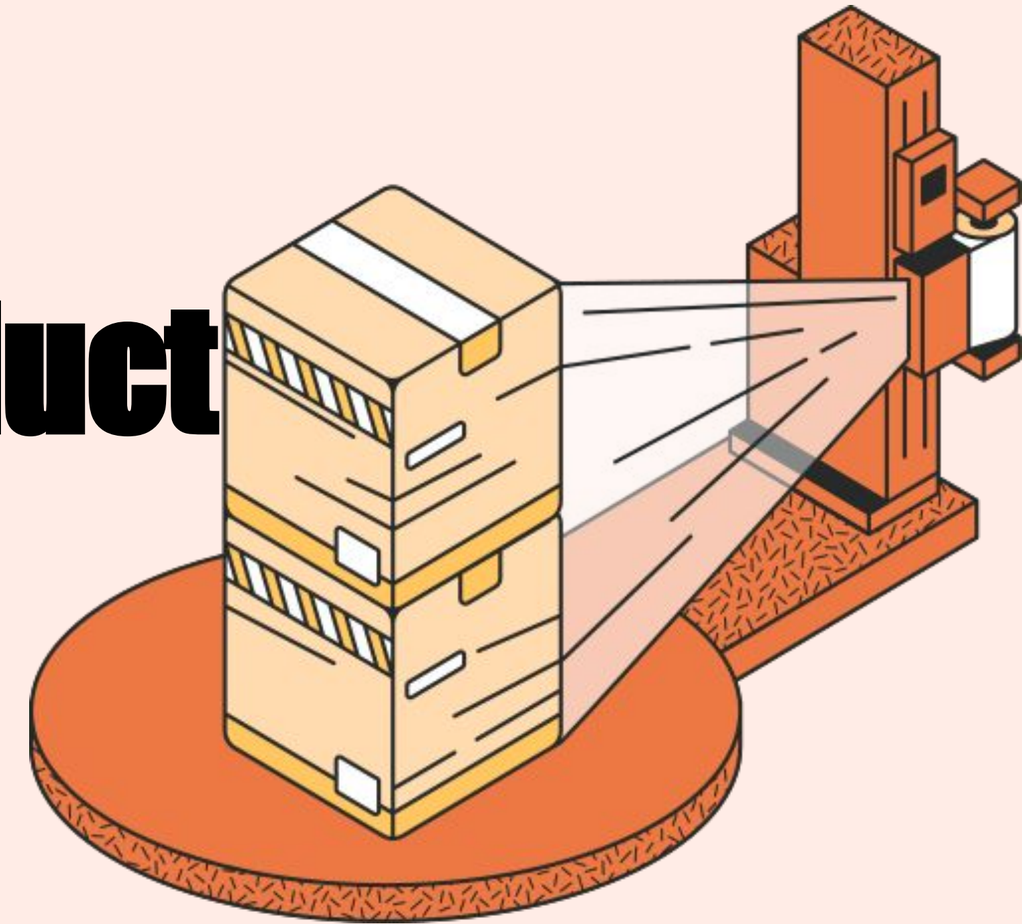
Customer 12607 cancelled all the orders placed (i.e. 100% of total sales)

80,000 units of PAPER CRAFT , LITTLE BIRDIE cancelled in Dec'11



Market Basket Analysis

Level I - Product



Creating a Sparse Matrix of Each Basket

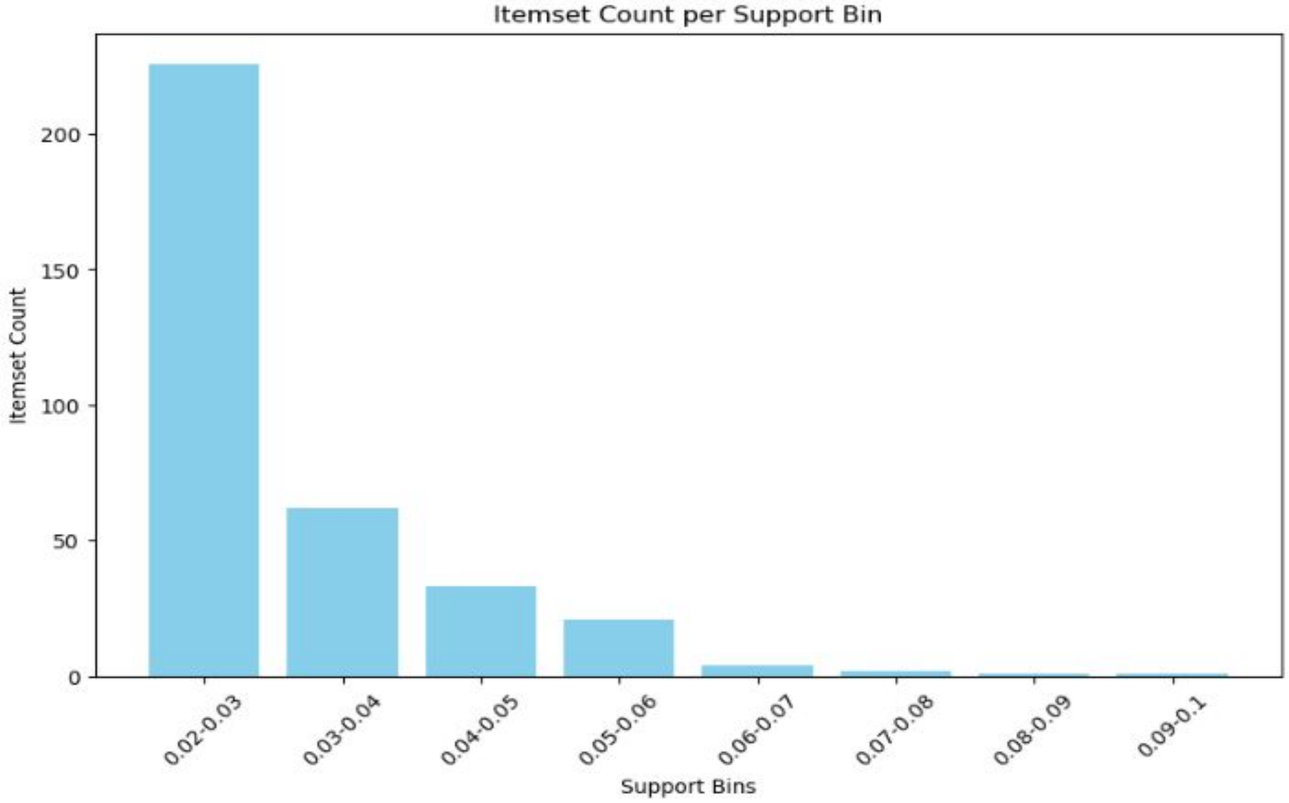
4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	DOLLY GIRL BEAKER	I LOVE LONDON MINI BACKPACK	I LOVE LONDON MINI RUCKSACK	NINE DRAWER OFFICE TIDY	OVAL WALL MIRROR DIAMANTE	RED SPOT GIFT BAG LARGE	SET 2 TEA TOWELS I LOVE LONDON	SPACEBOY BABY GIFT SET
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
...	...	...	...	...	...	...	...	...	...
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Overview of Product Scope

	support	itemsets
114	0.109656	(WHITE HANGING HEART T-LIGHT HOLDER)
46	0.101504	(JUMBO BAG RED RETROSPOT)
91	0.096507	(REGENCY CAKESTAND 3 TIER)
77	0.081805	(PARTY BUNTING)
63	0.075885	(LUNCH BAG RED RETROSPOT)
6	0.070597	(ASSORTED COLOUR BIRD ORNAMENT)
99	0.067200	(SET OF 3 CAKE TINS PANTRY DESIGN)
72	0.064047	(PACK OF 72 RETROSPOT CAKE CASES)
57	0.061766	(LUNCH BAG BLACK SKULL.)
69	0.060602	(NATURAL SLATE HEART CHALKBOARD)
44	0.059098	(JUMBO BAG PINK POLKADOT)
30	0.058273	(HEART OF WICKER SMALL)
55	0.057448	(JUMBO STORAGE BAG SUKI)
53	0.057011	(JUMBO SHOPPER VINTAGE RED PAISLEY)
37	0.056380	(JAM MAKING SET PRINTED)

Product scope represents the proportion of baskets containing a product.

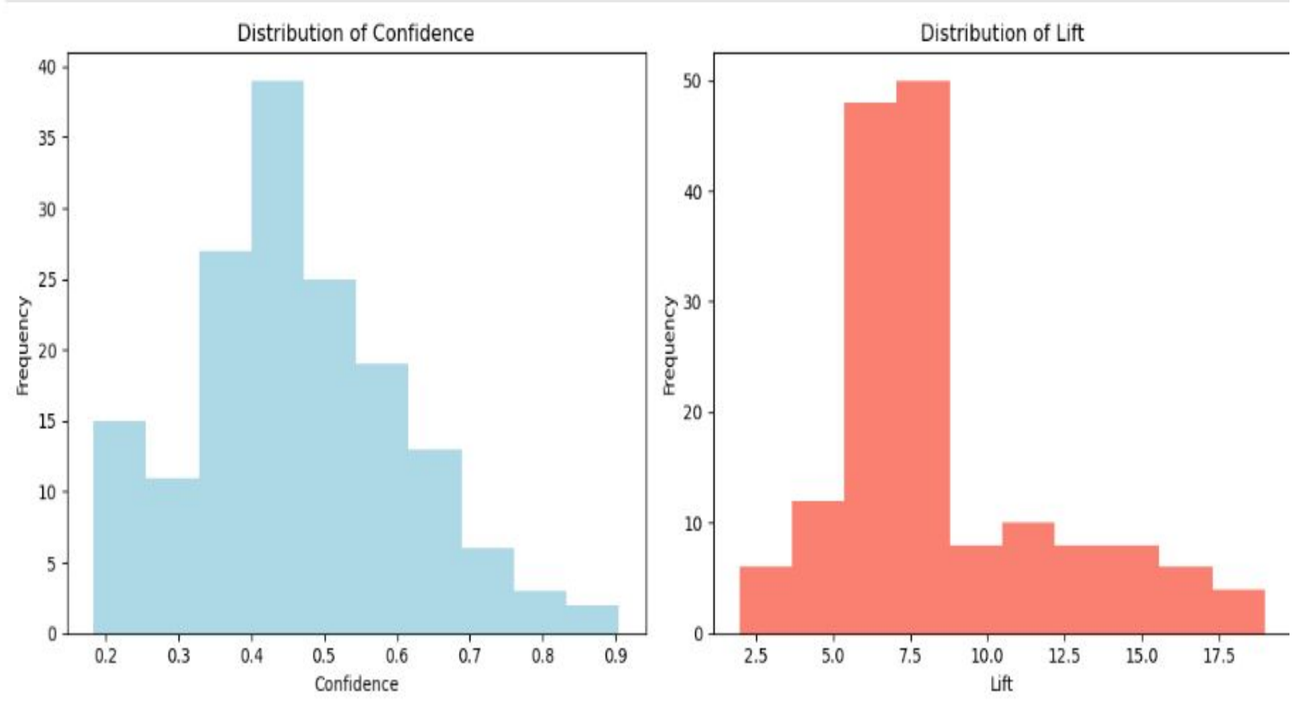
	Support Bin	Itemset Count
0	0.02-0.03	226
1	0.03-0.04	62
2	0.04-0.05	33
3	0.05-0.06	21
4	0.06-0.07	4
5	0.07-0.08	2
6	0.08-0.09	1
7	0.09-0.1	1

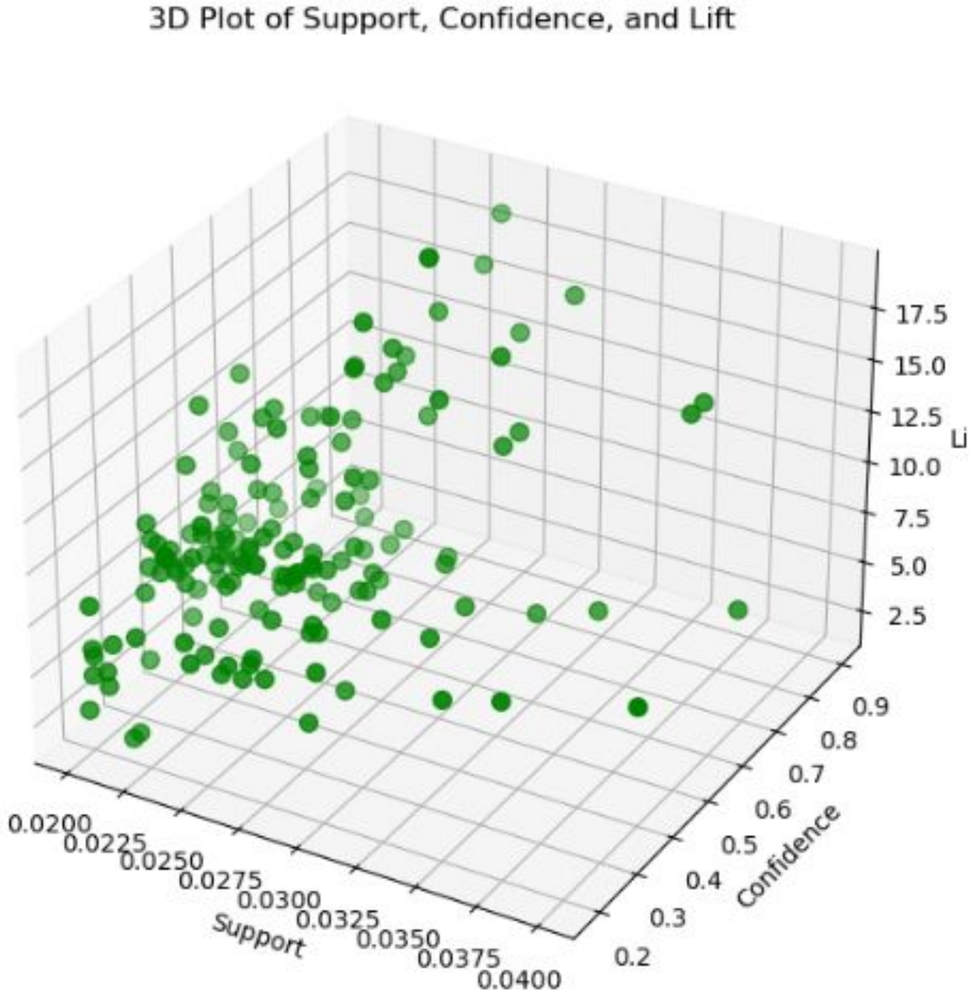


Experimenting with the Scope Threshold (Min Support 0.02)

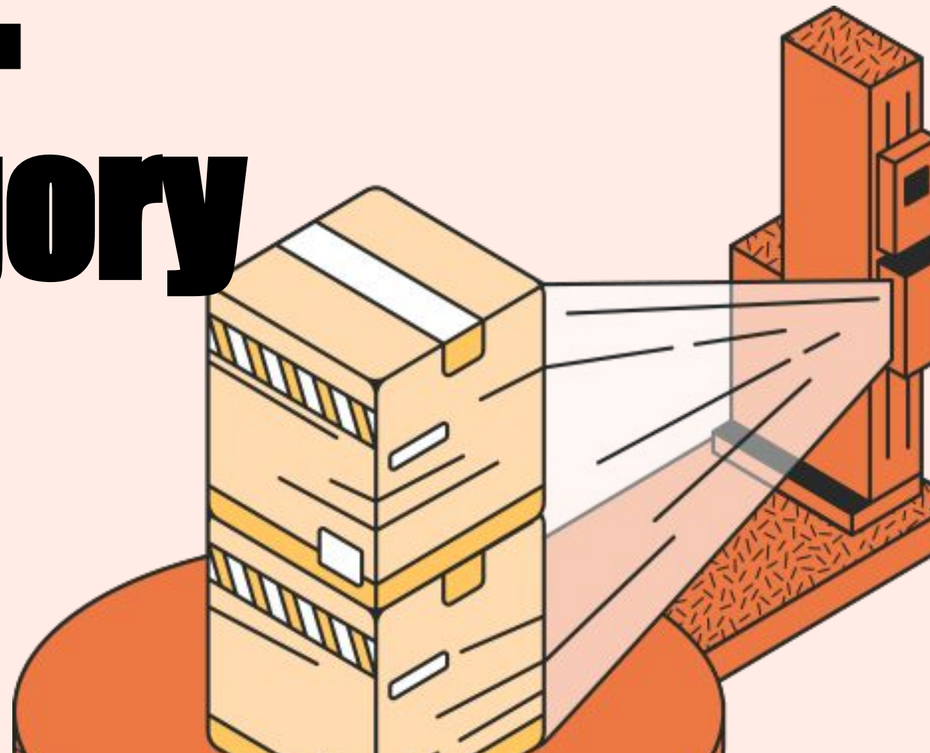
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs metric
(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.05	0.04	0.03	0.62	16.78	0.03	2.56	0.99
(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.04	0.05	0.03	0.83	16.78	0.03	5.48	0.98
(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.04	0.04	0.03	0.72	16.26	0.02	3.42	0.97
(GARDENERS KNEELING PAD KEEP CALM)	(GARDENERS KNEELING PAD CUP OF TEA)	0.04	0.04	0.03	0.6	16.26	0.02	2.4	0.98
(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.05	0.04	0.03	0.56	15.12	0.03	2.2	0.98

Insights from Product Level Analysis





Market Basket Analysis: Level II - Category



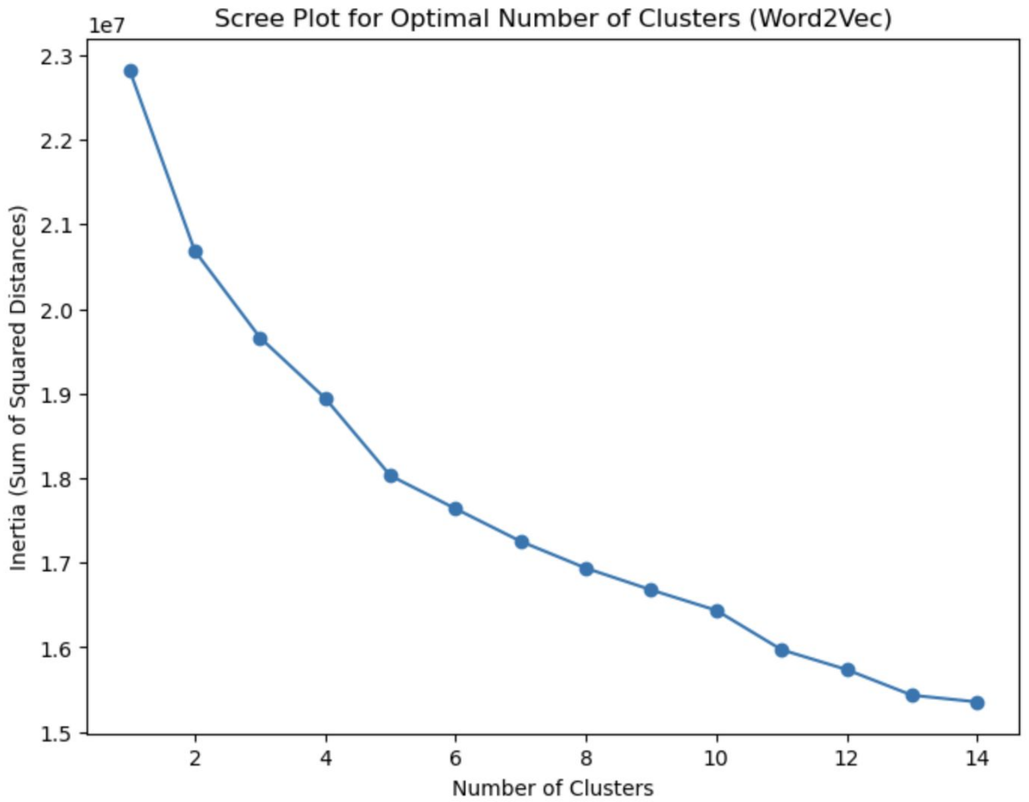
How We Categorized Products (Part 1)

We first tried a more complicated method:

- Used Word2Vec to vectorize each product description
- K-means clustering on description vectors

Challenges:

- No clear indication of how many clusters to use
- Manual interpretation of clusters required

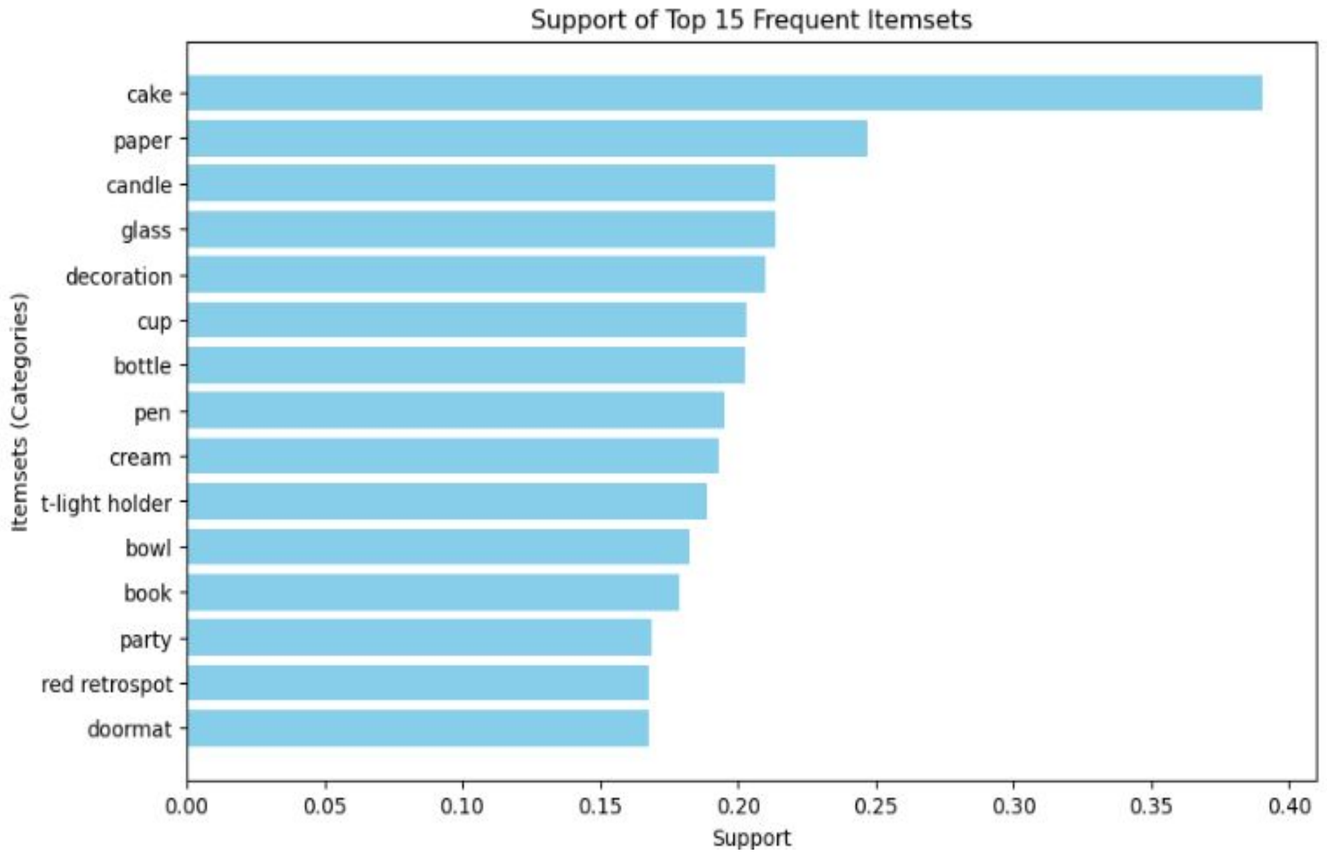


Scree plot from clustering based on Word2Vec

How We Categorized Products: Simpler is Better

Best method of product categorization was…

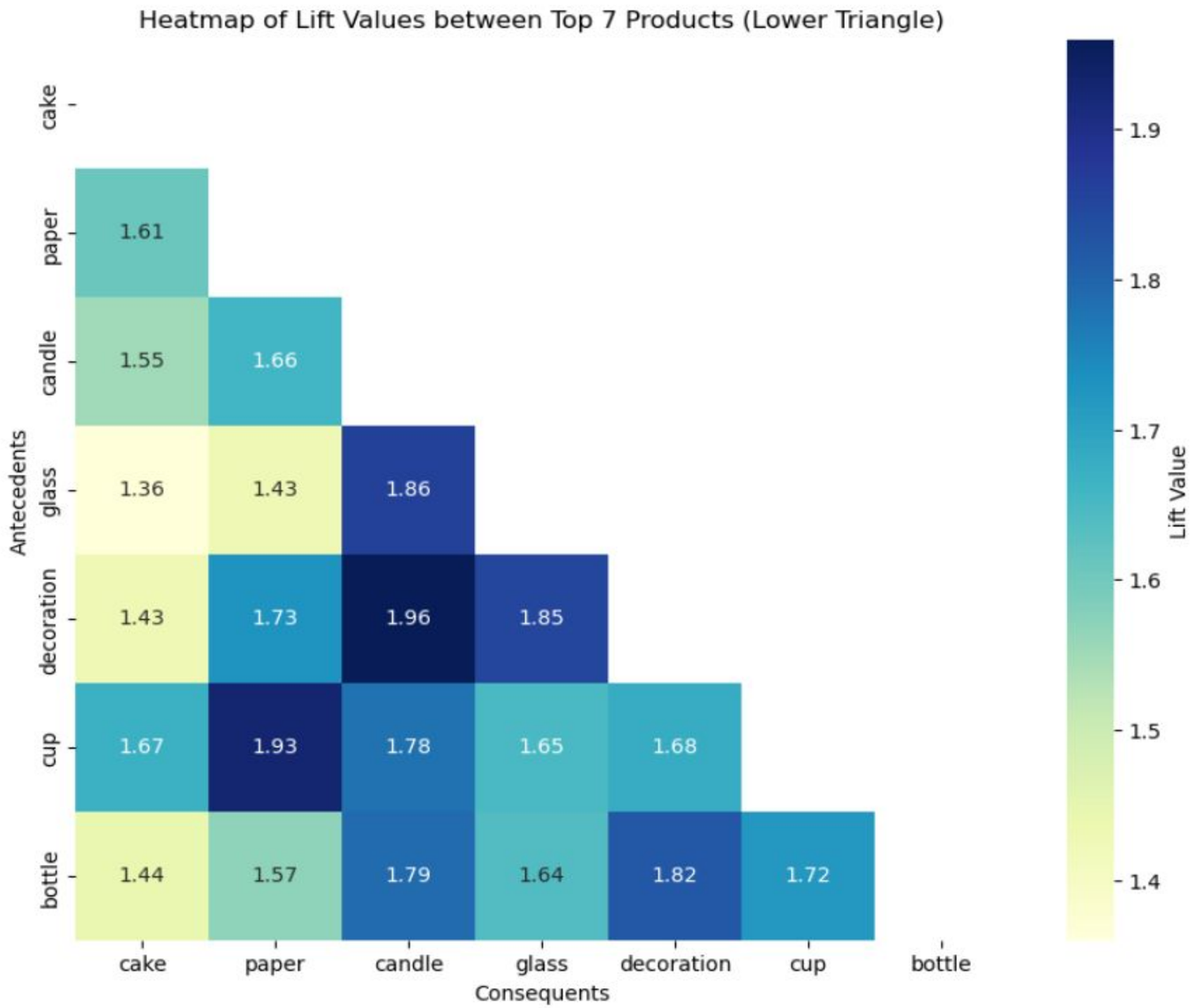
Extracting a list of keywords (cake, paper, bag, etc) and matching to product descriptions.



Results from Category Level Analysis

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(paint)	(decoration)	0.09	0.21	0.05	0.58	2.77	0.03	1.88	0.7
(bowl)	(cutlery)	0.18	0.11	0.05	0.29	2.55	0.03	1.25	0.74
(tissue)	(pen)	0.11	0.2	0.05	0.48	2.43	0.03	1.53	0.66
(cup)	(bowl)	0.2	0.18	0.09	0.44	2.42	0.05	1.46	0.74
(cup)	(cutlery)	0.2	0.11	0.06	0.27	2.41	0.03	1.22	0.73
(book)	(pen)	0.18	0.2	0.08	0.45	2.29	0.05	1.46	0.69
(book)	(flower)	0.18	0.15	0.06	0.33	2.25	0.03	1.28	0.68
(metal sign)	(frame)	0.16	0.16	0.06	0.37	2.24	0.03	1.32	0.66
(bin)	(frame)	0.15	0.16	0.06	0.37	2.23	0.03	1.32	0.65
(metal sign)	(doormat)	0.16	0.17	0.06	0.36	2.16	0.03	1.31	0.64
(pink polkadot)	(bowl)	0.13	0.18	0.05	0.39	2.14	0.03	1.34	0.61
(flower)	(decoration)	0.15	0.21	0.07	0.45	2.14	0.04	1.44	0.63
(book)	(bowl)	0.18	0.18	0.07	0.38	2.1	0.04	1.33	0.64
(metal sign)	(party)	0.16	0.17	0.06	0.35	2.1	0.03	1.29	0.63
(cream)	(flower)	0.19	0.15	0.06	0.31	2.1	0.03	1.24	0.65
(t-light holder)	(frame)	0.19	0.16	0.06	0.34	2.1	0.03	1.27	0.64
(gift)	(pen)	0.13	0.2	0.05	0.41	2.1	0.03	1.36	0.6

Heatmap of Lift Values of Top 7 products



Results from n:1 Market Basket Analysis

Multi Product Mapping

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
772	(cake, red retrospot)	(pink polkadot)	0.10	0.13	0.05	0.55	4.28	0.04	1.94	0.85
550	(cup, paper)	(bowl)	0.10	0.18	0.05	0.53	2.92	0.03	1.75	0.73
514	(cup, cake)	(bowl)	0.13	0.18	0.07	0.52	2.83	0.04	1.69	0.75
532	(metal sign, cake)	(bowl)	0.10	0.18	0.05	0.50	2.76	0.03	1.64	0.71
424	(book, cake)	(bowl)	0.12	0.18	0.06	0.49	2.67	0.03	1.59	0.71
574	(candle, cake)	(flower)	0.13	0.15	0.05	0.39	2.59	0.03	1.39	0.71
460	(book, cake)	(pen)	0.12	0.20	0.06	0.50	2.55	0.03	1.60	0.69
670	(cup, cake)	(party)	0.13	0.17	0.06	0.42	2.49	0.03	1.43	0.69
508	(cake, cream)	(bowl)	0.12	0.18	0.05	0.45	2.48	0.03	1.49	0.68
502	(candle, cake)	(bowl)	0.13	0.18	0.06	0.45	2.48	0.04	1.49	0.69
706	(cake, t-light holder)	(decoration)	0.10	0.21	0.05	0.52	2.47	0.03	1.64	0.66

Insights

High Lift Pairings: Strong product associations (e.g., "cake, red retrospot" with "pink polkadot") suggest valuable cross-selling opportunities.

Multi-Product Combinations: Items like "cup, cake" and "bowl" are frequently bought together, ideal for bundled promotions.

Category-Based Grouping: Related items (e.g., kitchen and decor) show high association, supporting category-focused marketing strategies.

Heatmap Insights: High lift values in top products (e.g., "cake", "candle") guide targeted promotions.

Store Layout Optimization: Position high-association items together for better customer convenience and increased sales.

Cross-Selling Potential: High association pairs can drive additional sales by leveraging natural buying patterns.

Recommendations

Cross-Promotion and Bundling

Create bundled offers for frequently purchased product pairs and apply bundle discounts to encourage customers to buy multiple items together, enhancing basket size and overall sales.

- Bundle pairs like Paint & Decorations.
- Offer discounts (e.g., 10% off for 2 items)
- Test bundle combos to find top-sellers.

Product Placement Optimization

Feature “Frequently Bought Together” items prominently on product pages to promote add-on purchases, leveraging customer buying patterns to increase the likelihood of multi-item transactions.

- Place items together on relevant pages.
- Choose best location on the webpage
- A/B test layouts for best results.

Inventory Planning and Reordering

Maintain balanced stock levels for popular product pairs by automating reorders when one item in a pair is low, ensuring both products remain available to meet customer demand and prevent stockouts.

- Set low-stock alerts for popular pairs.
- Sync restocks to avoid shortages.
- Monitor peak seasons for demand spikes.

(dls) Honorable Mentions!

- **Find the Valuable Customers for Targeting Activity:** The lack of direct response data limits our ability to measure the effectiveness of targeted marketing campaigns for each RFM segment.
- **Prediction of Purchase Qty Based On Invoice Date and Price:** Despite feature engineering and log transformation, the features were not predictive of the number of units purchased (using linear regression and random forest).
- **Market Basket Analysis at a Higher-Level Grouping:** Due to data sparsity and insufficient category definitions, many products could not be categorized effectively, limiting the ability to draw meaningful, higher-level insights for strategic product bundling and cross-category promotions.

Future Scope

- **Customer Segmentation Using Cluster Analysis**
  - *Need*: Demographic and psychographic data on customers.
  - *Action*: Use cluster analysis to identify customer segments based on purchasing behavior for targeted marketing.

- **CLV Modeling for High-Value Customer Targeting**
  - *Need*: Long-term purchase and engagement history.
  - *Action*: Prioritize marketing resources towards high-value segments and inform loyalty program design using CLV analysis.

- **A/B Testing for Campaign Effectiveness**
  - *Need*: Historical data on campaign types and customer responses.
  - *Action*: Evaluate the effectiveness of different marketing campaigns using A/B tests to optimize strategies based on measured customer responses.

- **Time-Series Analysis for Purchase Trends and Seasonality**:
  - *Need*: Extended multi-year dataset to capture seasonality.
  - *Action*: Identify demand patterns using time-series analysis, helping adjust inventory and marketing strategies.

Thank You

Appendix

Lift Values at Minimum support of 0.02

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
148	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	(PINK REGENCY TEACUP AND SAUCER)	0.037263	0.037166	0.026298	0.705729	18.988353	0.024913	3.271930	0.984004
153	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENC...	0.037166	0.037263	0.026298	0.707572	18.988353	0.024913	3.292215	0.983904
150	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	(GREEN REGENCY TEACUP AND SAUCER)	0.029064	0.049248	0.026298	0.904841	18.373184	0.024867	9.991237	0.973877
151	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	0.049248	0.029064	0.026298	0.533990	18.373184	0.024867	2.083511	0.994553

RFM Analysis

Segmentation of R, F and M into 10 buckets

Identifying top customers by filtering for R ≥ 8, F ≥ 8, M ≥ 8

Identified 347/4380 unique customers as our “probable” best customer segment

	A	B	C	D	E	F	G	H
1	Row Labels	Max. of InvoiceDate	Count of InvoiceNo	Sum of UnitPrice	Recency	Frequency	Monetary	RFM Score
3	12347	07/12/2011	182	481.21	9	8	8	988
14	12359	02/12/2011	254	2225.11	8	9	9	899
17	12362	06/12/2011	274	1083.29	9	9	9	999
32	12381	05/12/2011	91	417.04	9	7	8	978
36	12388	24/11/2011	100	277.77	7	7	7	777
41	12395	24/11/2011	159	629.59	7	8	8	788
58	12417	06/12/2011	198	1074.23	9	8	9	989
63	12423	09/12/2011	126	365.05	9	8	7	987
69	12429	30/11/2011	97	352.18	8	7	7	877
73	12433	09/12/2011	420	920.95	9	9	9	999
77	12437	08/12/2011	201	809.45	9	8	9	989
78	12438	25/11/2011	98	381.7	7	7	8	778
81	12444	18/11/2011	177	1119.13	7	8	9	789
88	12451	29/11/2011	355	868.18	8	9	9	899
102	12471	07/12/2011	531	2266.87	9	9	9	999
105	12474	22/11/2011	457	1331.7	7	9	9	799
107	12476	08/12/2011	264	1306.93	9	9	9	999
117	12490	04/12/2011	245	659.75	8	9	8	898