School of Informatics & IT
TEMASEK POLYTECHNIC

DIPLOMA IN BIG DATA & ANALYTICS

DATA WAREHOUSING AND BUSINESS INTELLIGENCE (CDA2C01)

AY 2023 APRIL SEMESTER

Project – Individual Report

Practical Class: P03

Tutor: Percy Wong

Submitted by Javen Lai Le Yu

Student Name: Javen Lai Le Yu

Matric Number: 2202934B

Data Warehousing and Business Intelligence (CDA2C01)

AY 2023 APRIL SEMESTER

Project – Individual Report

Student Name (Matric Number): JAVEN LAI LE YU

Practical Group: P03

Tutor: PERCY WONG

Submission Date: 24/8/2023

Declaration of Originality

I am the originator of this work and I have appropriately acknowledged all other original sources used as my reference for this work.

I understand that Plagiarism is the act of taking and using the whole or any part of another person's work, including work generated by AI, and presenting it as my own.

I understand that Plagiarism is an academic offence and if I am found to have committed or abetted the offence of plagiarism in relation to this submitted work, disciplinary action will be enforced.

Content covered:

Introduction

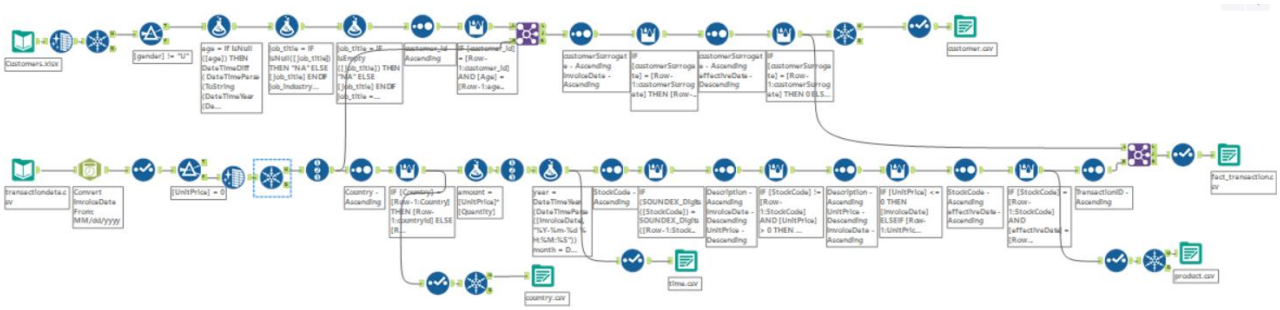
This project aims to **understand the business performance in Spain** by completing the following tasks:

1. Identify the **best and worst-performing products**.
2. Analyze the **sales trends of our products** across different time periods (day, week, month).
3. Identify any **missed opportunities**.
4. Examine the **variations in product sales across different countries**.
5. Identify seasonal or **periodic patterns in product sales**.
6. Identify **cross-selling and upselling opportunities** based on customer purchase patterns.

Finally, I will develop a data-driven **plan to increase sales performance** for our platform.

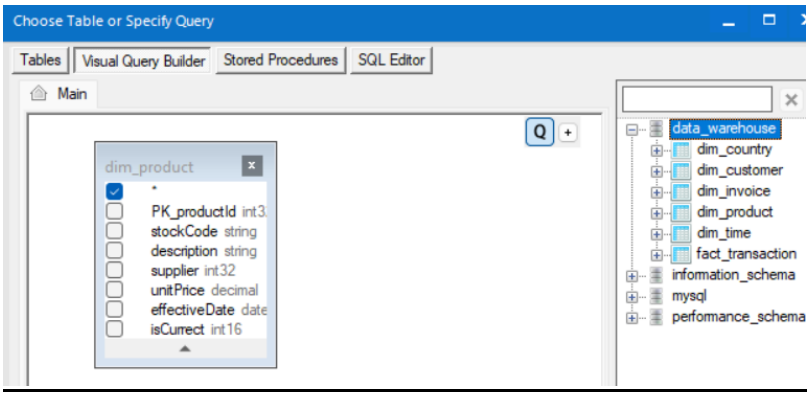
Data Warehouse

My team **extracted** the retail company's historical transactions and customer database. Then, we performed **transformations** to prepare it for being **loaded** into the Data Warehouse we designed.



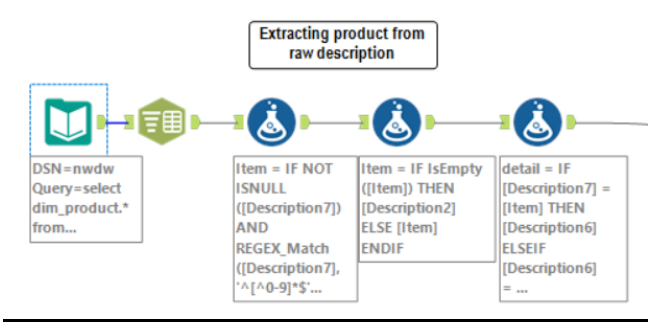
The datasets underwent thorough **cleaning**, such as **formatting date** and **removing duplicated and improbable records** (e.g. age 123 with active status). **Missing values** in age, job title, and industry were **imputed**. **New columns**, such as **Surrogate Keys, Amount** (Quantity*UnitPrice), and **Slow-Changing Dimension** (SCD) **constraints**, were also generated. Finally, we **formatted the datatypes** using Select Node and **filtered** out columns and rows (duplicated records using Unique Node) not required for each dimension table. Then, we created a **CSV file for each dimension and loaded** these files **into the Data Warehouse using an SQL script**.

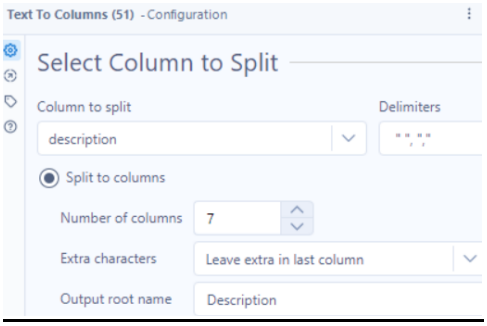
Data Extraction



To analyze the business transactions, I **loaded data from** each table of my **Data Warehouse** individually **using Input Data Node**. I will **connect tables using Joiner Node** when required as combining everything at the start causes slow runtime due to an inefficient workflow.

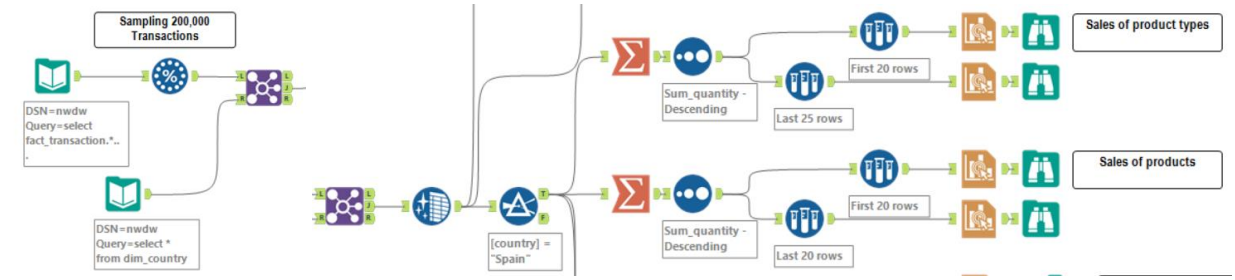
Preprocessing Product description





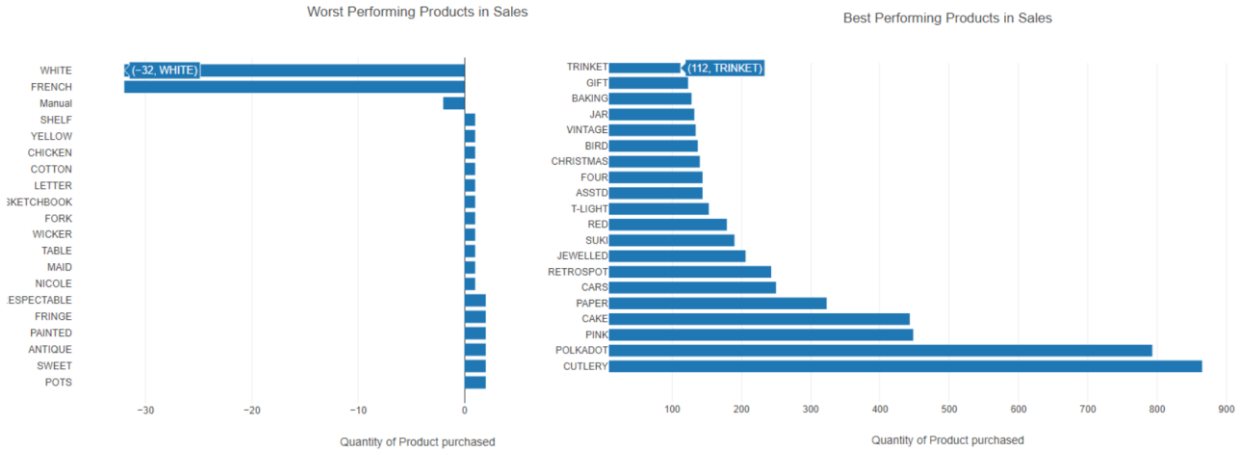
I **splitted** the lengthy description into 7 parts and **extracted** the **last word** that contains no numbers or lowercase **of the description to be the product type** (item), and the **second last word to be the product** (detail).

Randomly Sampling 200,000 records, using 22029 as the seed:



After **sampling** 200,000 transaction records, I **joined** these records with the product's workflow to **obtain the item and detail**. Then, I **cleansed** the country column of whitespaces and **filtered** out records that are not Spain. Using **Summarize** Node, I derived the total quantity for each product type and product. I **sorted** by descending and sampled the first and last 20 rows to find the best and worst performer using interactive charts directly.

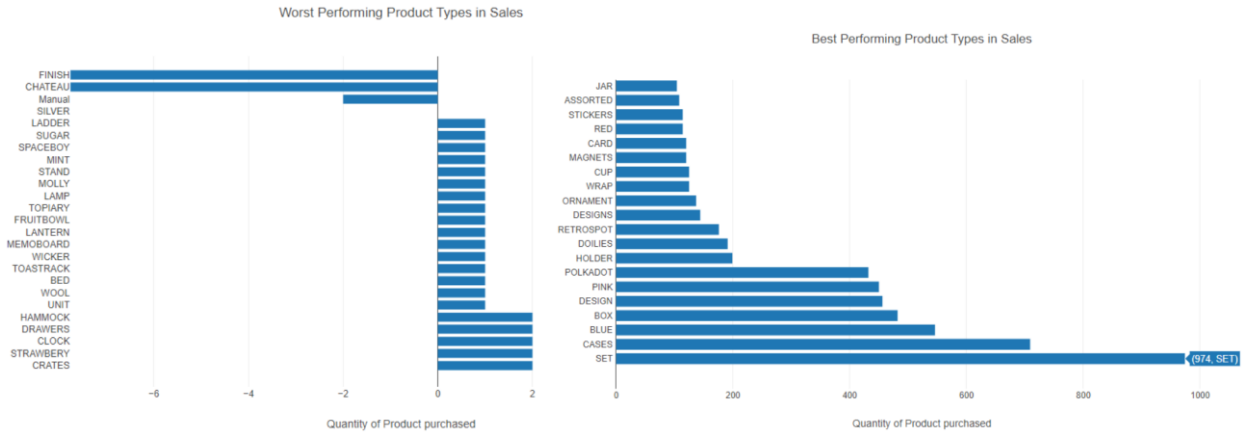
Identifying best and worst-performing products



Best performing products: **Kitchenware** like Cutlery, Jar, Cake products, and Baking related products. **Products with designs** such as Polkadot, Pink, Red, and Jewelled. Products under **Retrospot** brand. **Vintage** products.

Worst performing products: **Large furnitures** like a shelf or table. **Antiques**, pots, sketchbooks, and letters. **White and French** related products are making a loss.

Identifying best and worst-performing product types



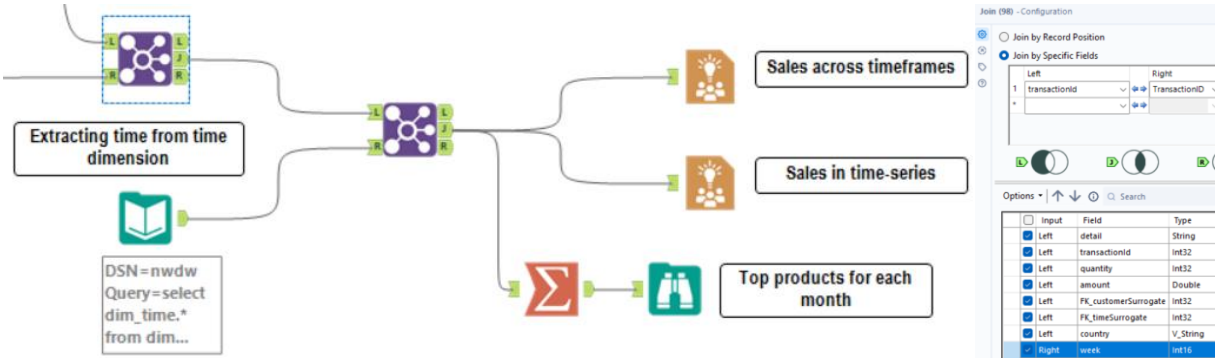
Best performing product types: Customers tend to **buy items in a set**. They often buy **case, box, wrap, or ornament** related products. They **prefer blue, polkadot, retrospot, and design** items.

Worst performing product types: Customers **rarely purchase appliances or furniture** like clocks, lamps, ladders, drawers, hammocks, or beds.

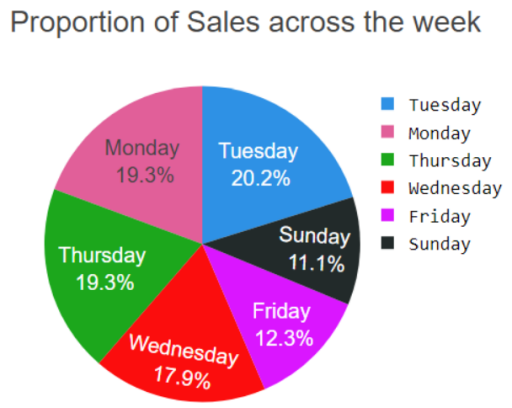
Sales trend of products across different time periods

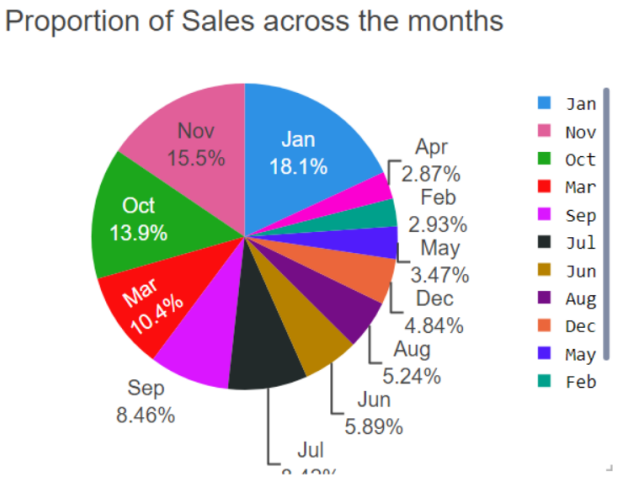
I created 'week' to analyze sales in time series. I reused my group's workflow to ensure the transactionId created from RecordID is identical to the existing transactions' ID from my data warehouse.

Then, I **joined this workflow with the filtered Spain transactions** using transactionId as the join constraint:



To obtain the time features, I extracted them from the time dimension. I joined using timeSurrogate as the join constraint. Now that we have obtained the required time features for each transaction, we can proceed to analyze it.





Sales refer to the quantity of goods sold. **Sales are the poorest during weekends**, and there are **no sales made on Saturday for the 2 years**. This likely indicates that the **platform is closed on Saturday. Sales are best during January, November, and October**.

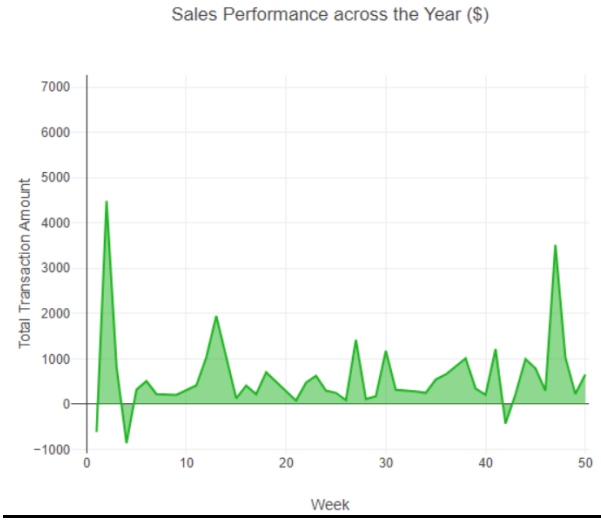
month	Mode_detail	Mode_Item	Mode_supplier
Jan	LUNCH	BOX	235
Apr	PAPER	SET	82
Jun	RETROSPOT	PINK	293
Oct	CAKE	DECORATION	228
Mar	CAKE	CASES	180
Nov	CHRISTMAS	CHRISTMAS	230
Dec	VINTAGE	BAG	46
May	LUNCH	BOX	247
Aug	OVEN	DESIGN	302
Sep	CAKE	CASES	297
Feb	3	HOLDER	215
Jul	VINTAGE	CASES	228

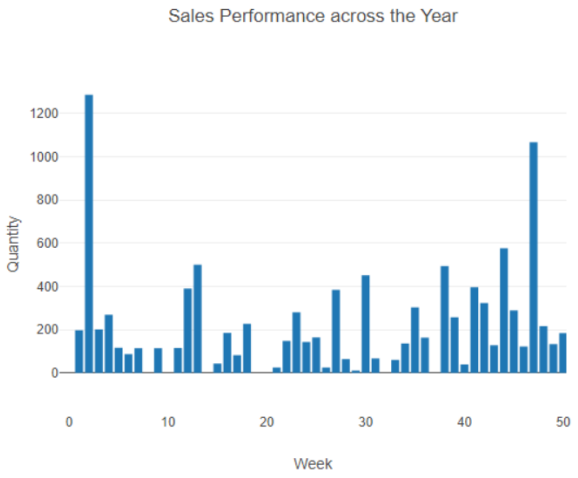
Trend of Sales for each month

Customers buy **Lunch Boxes or Bags** at the start of the year possibly **for packing food to bring to work or school**. For October, March, and September, customers tend to buy **cake-related products**. Could they be **celebrating events** within those months? Or is **Cake baking popular in Spain?**

Lastly, Christmas products and product types are the most popular products during November, indicating that **customers use our platform to purchase Christmas products** like costumes and decorations to prepare for Christmas.

Periodic pattern in product sales (time-series)





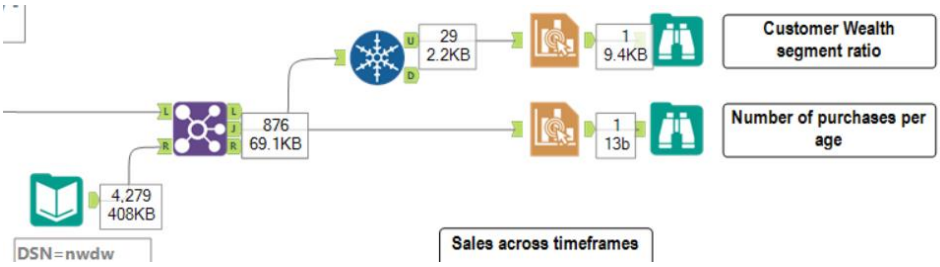
Sales peak in the 2nd week, likely due to Customers **buying lunch-related products like lunch boxes and lunch bags to be used for the new year**, to **bring packaged food to their workplace** or school, or a picnic. Sales also peak at the **47th week, likely due to customers buying a lot of Christmas products before the Christmas** month.

There are certain **weeks with 0 sales (no products sold) for the past 2 years**. **Was our platform down for the same exact weeks for the past 2 years?** Because sales for the week prior to the 0 sales weeks seem to be performing fine. Another surprising insight is that sales of goods do not always translate to profit, as some weeks had products being sold, but still **made a loss in transaction amount**.

stockCode	supplier	unitPrice	effectiveDate	product_flag	Item	detail	transactionId	quantity
82483	486	5.95	2010-12-01	0	FINISH	WHITE	532,302	-32
23245	297	10.79	2011-06-09	0	TINS	CAKE	532,298	-16
82486	486	8.15	2011-09-05	0	CABINET	WOOD	532,303	-12
22173	190	2.75	2011-10-07	0	CHATEAU	FRENCH	532,290	-32
23072	280	6.5	2011-10-14	0	SILVER	BOX	532,297	-48
82482	486	2.55	2010-12-02	0	FINISH	WHITE	532,301	-72
22553	228	3.36	2010-12-01	0	SKULLS	TIN	532,928	-12
21843	158	10.95	2010-12-01	0	STAND	CAKE	532,926	-2
M	633	1,715.85	2011-01-27	0	Manual	Manual	526,146	-1
84817	553	2.1	2010-12-19	0	PLATE	DECORATIVE	534,053	-4
22059	178	1.49	2010-12-06	0	MUG	DESIGN	534,049	-2

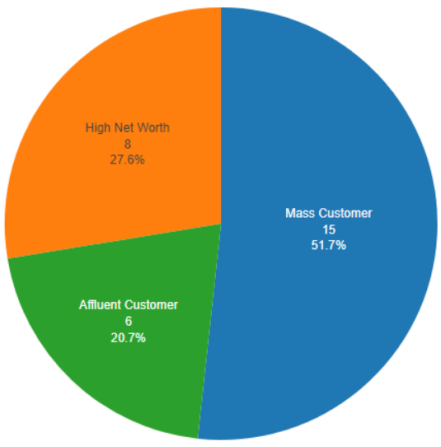
This is due to refunds of products. This indicates that **some of our products are of an unsatisfactory quality** which led to customers refunding.

Understanding our Customers

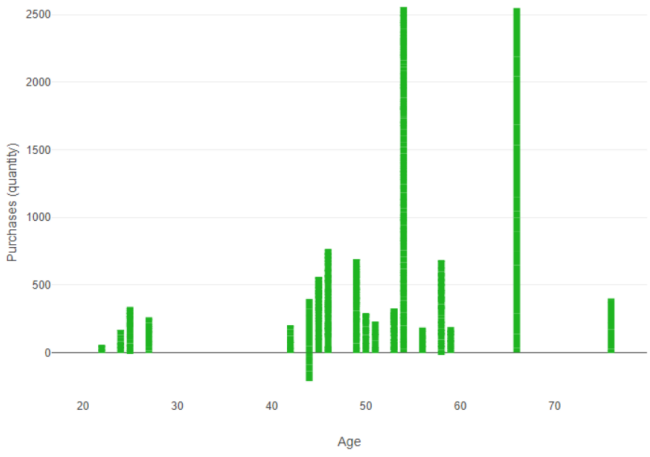


I used unique node to filter out only 1 record for each customerId.

Customer Type in Spain



Number of purchases made per age group

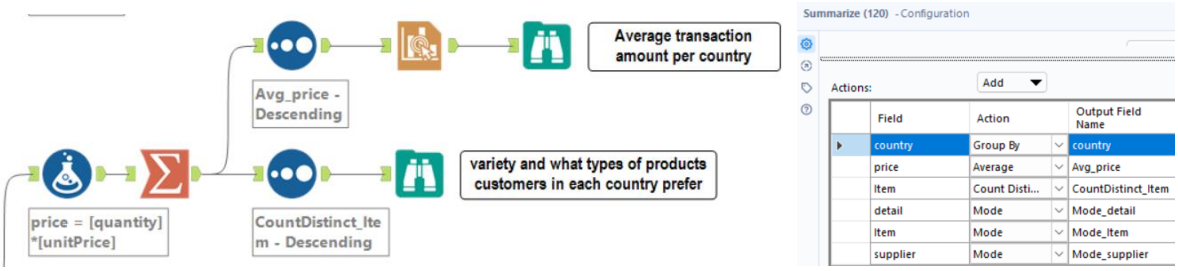


Spain only has **29 Customers**, and **most of them are mass customers**. **Most purchases come from customers aged 42-60**, and our **regular customers are ages 54 and 66**. Spain has no customers aged 28 to 41, meaning there are **no middle-aged working adults** using our platform to make purchases, which is a missed opportunity as this age group has spending power as they have worked for years in their life and should have adequate savings to splurge on shopping. I also discovered that **refunds were mainly made by customers aged 44**.

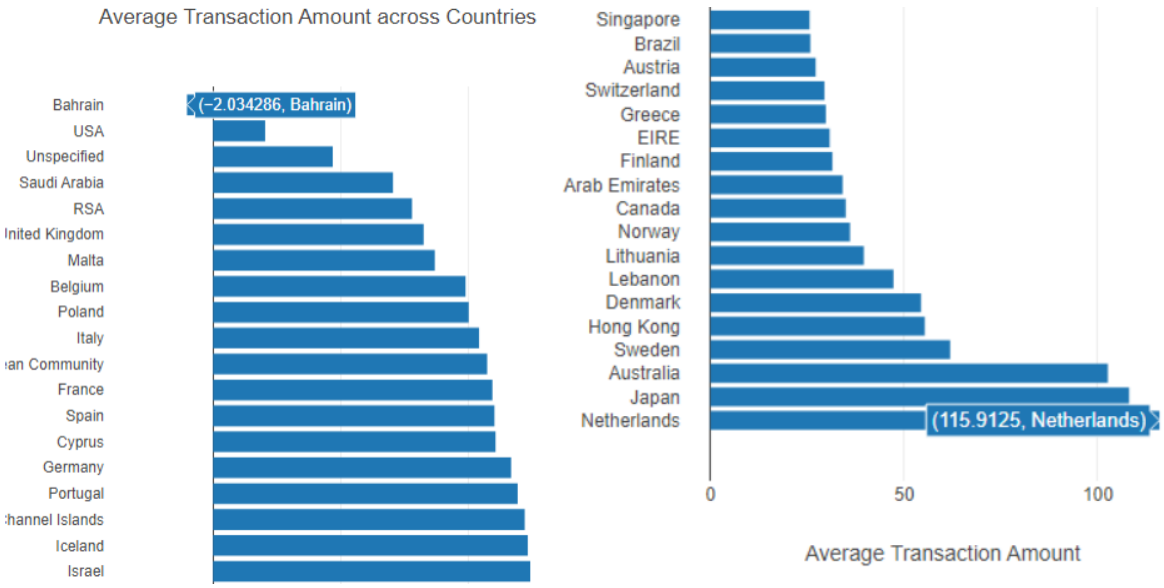
Identifying missed opportunities

1. Weekend performance is extremely poor despite it being the period where Customers have free time to do whatever they like – such as online shopping! We have **failed to attract Customers during weekends**.
2. The platform is likely closed on Saturdays. The **platform should open up on Saturdays** to increase sales, as customers would be free during weekends to shop using our platform.
3. **Downtime might have disrupted sales** by preventing customers from using our Platform. We **could have grossed higher sales** from purchases by customers **if our platform was running** during those periods.
4. Customers are **not fully utilizing our platform for celebrating events** except for Christmas. The ideal outcome we could strive to achieve is for the top-performing product of each month to be event-related products: e.g. Valentine's, Easter, or Halloween.
5. We could have made more revenue if we **ensured our products were of satisfactory standards** so that fewer customers would make refunds. We should **understand customers aged 44 to find out what causes them to refund products**.
6. We have **failed to reach the market of middle-aged adults,** as there are no purchases from Customers aged 28-41.

Variations in product sales across different countries



I group by country, and then summarized each measure by taking either the mean, count, or mode.



An **average Spain Customer spends around $22 per transaction**. **Bahrain is making a loss** on average. This could be due to **expenses overshadowing profits** and/or **extensive refunds** made by unsatisfied customers. Customers from the **Netherlands spend the most per transaction on average**, as each transaction they make is around $116. This could indicate that **they buy in bulk, buy many products at once, or buy expensive products**.

Top Product, Product Type, and Supplier for each country

country	CountDistinct_Item	Mode_detail	Mode_Item	Mode_supplier	Avg_amount
United Kingdom	869	VINTAGE	DESIGN	293	16.475647
EIRE	468	CAKE	DESIGN	290	30.76461
France	397	POSTAGE	DESIGN	635	21.853418
Germany	391	POSTAGE	DESIGN	635	23.316366
Spain	256	CAKE	DESIGN	228	22.012705
Switzerland	232	PAPER	DESIGN	228	29.423684
Netherlands	219	CAKE	DESIGN	235	115.912537
Belgium	199	CAKE	DESIGN	635	19.753707
Portugal	186	VINTAGE	DESIGN	186	23.830441
Norway	162	CAKE	DESIGN	228	36.0353
Australia	160	VINTAGE	DESIGN	302	102.588218
Cyprus	140	AND	DESIGN	290	22.098917
Italy	139	BAG	DESIGN	257	20.811455
Channel Islands	128	CAKE	DESIGN	269	24.380432
Finland	121	POSTAGE	POSTAGE	635	31.488838
Unspecified	99	CAKE	CASES	72	9.367375
Austria	92	POLKADOT	POSTAGE	635	27.158811
Denmark	82	LUNCH	BOX	235	54.340709
Poland	78	HANGING	DESIGN	235	20.00073
Sweden	77	CAKE	CASES	635	61.881195
Israel	73	TINS	BAG	98	25.428707
Japan	68	PAPER	DESIGN	95	108.076667
USA	61	CAKE	CASES	86	4.077438
Hong Kong	57	POLKADOT	BOWL	42	55.343578
Singapore	54	RED	RETROSPOT	43	25.58375
Iceland	49	BAKELIKE	CASES	245	24.624722
Canada	46	EXERCISE	BOOKS	264	34.920909
Greece	43	DOLLY	ASSORTED	265	29.795455
Malta	31	FRAME	CORNICE	249	17.346889

```
1 •  SELECT DISTINCT description, supplier FROM data_warehouse.dim_product
2    WHERE supplier IN (635, 228, 235)
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content:

description	supplier
HOLIDAY FUN LUDO	228
PLASTERS IN TIN SPACEBOY	228
PLASTERS IN TIN SKULLS	228
PLASTERS IN TIN WOODLAND ANIMALS	228
PLASTERS IN TIN STRONGMAN	228
PLASTERS IN TIN CIRCUS PARADE	228
PLASTERS IN TIN VINTAGE PAISLEY	228
CLOTHES PEGS RETROSPOT PACK 24	228
SEASIDE FLYING DISC	228
4 TRADITIONAL SPINNING TOPS	235
TRADITIONAL KNITTING NANCY	235
BOX OF VINTAGE ALPHABET BLOCKS	235
BOX OF VINTAGE JIGSAW BLOCKS	235
IVORY KITCHEN SCALES	235
RED KITCHEN SCALES	235
BLACK KITCHEN SCALES	235
MINT KITCHEN SCALES	235
PICNIC BOXES SET OF 3 RETROSPOT	235
SPACEBOY LUNCH BOX	235
POSTAGE	635

Customers from **many countries like design items**. **Cake, Vintage and Polkadot products are popular in several countries**. For **certain countries, the most common transactions are postages**. The market on our **platform isn't dominated by any supplier**, but some notable suppliers are 635, 228, and 235 (Top Suppliers for multiple countries). These suppliers sell **Plasters** in a tin, **kitchen scales**, **lunch, and picnic boxes**. 635 is the supplier of postage service. An interesting insight is that these top **suppliers sell many types/patterns of the same product**. Plasters may also be a high-demand item in Spain as the top supplier of Spain sells mostly plasters.

```
1 •  SELECT p.supplier, p.description, p.unitprice, SUM(t.quantity) AS `Quantity Sold`
2        FROM dim_product p JOIN fact_transaction t ON p.PK_productId = t.FK_productId
3        WHERE supplier IN (228, 235)
4        GROUP BY p.PK_productId
5        ORDER BY description, `Quantity Sold` DESC
```

sult Grid | Filter Rows: | Export: | Wrap Cell Content:

supplier	description	unitprice	Quantity Sold
235	4 TRADITIONAL SPINNING TOPS	1.45	2481
235	4 TRADITIONAL SPINNING TOPS	2.51	1169
235	4 TRADITIONAL SPINNING TOPS	1.06	960
235	4 TRADITIONAL SPINNING TOPS	2.46	220
235	BLACK KITCHEN SCALES	8.50	306
235	BLACK KITCHEN SCALES	7.65	156
235	BLACK KITCHEN SCALES	16.63	12
235	BLACK KITCHEN SCALES	16.98	5
235	BOX OF VINTAGE ALPHABET BLOCKS	21.23	582
235	BOX OF VINTAGE ALPHABET BLOCKS	8.50	348
235	BOX OF VINTAGE ALPHABET BLOCKS	11.95	319
235	BOX OF VINTAGE ALPHABET BLOCKS	20.79	13
235	BOX OF VINTAGE ALPHABET BLOCKS	24.96	5
235	BOX OF VINTAGE JIGSAW BLOCKS	5.95	470
235	BOX OF VINTAGE JIGSAW BLOCKS	4.95	336
235	BOX OF VINTAGE JIGSAW BLOCKS	4.25	80
235	BOX OF VINTAGE JIGSAW BLOCKS	12.46	13
235	BOX OF VINTAGE JIGSAW BLOCKS	11.02	7
235	BOX OF VINTAGE JIGSAW BLOCKS	10.79	6
228	CLOTHES PEGS RETROSPOT PACK 24	2.51	3105
228	CLOTHES PEGS RETROSPOT PACK 24	1.65	2910
228	CLOTHES PEGS RETROSPOT PACK 24	1.45	2334
228	CLOTHES PEGS RETROSPOT PACK 24	3.29	142
228	CLOTHES PEGS RETROSPOT PACK 24	3.95	40
228	CLOTHES PEGS RETROSPOT PACK 24	2.46	33
228	HOLIDAY FUN LUDO	3.75	593
228	HOLIDAY FUN LUDO	3.39	288
228	HOLIDAY FUN LUDO	2.62	216
228	HOLIDAY FUN LUDO	7.46	28
228	HOLIDAY FUN LUDO	7.62	21

Are discounts effective?

Products for top-performing suppliers all had price changes. It is quite **evident that discounts** do in fact help boost sales.

This can be seen from BLACK KITCHEN SCALE, VINTAGE JIGSAW BLOCKS, and HOLIDAY FUN LUDO, which probably had a one-for-one deal as the discount version was half the price, and this marketing strategy was effective as **more quantity was sold when the price was slashed by half.**

For RETROSPOT's CLOTHES PEGS and SPINNING TOPS, the **slightly discounted versions performed better** than the full price; *most* **customers only bought when there was a discount**.

This proves that **giving slight discounts or one-for-one promotions can significantly increase sales** based on customers' behavior.

Upselling and Cross-Selling Opportunities for Spain Customers

Upselling:

Since **Customers in Spain make moderate purchases** (~$22), we can **make premium products more affordable** by giving small discounts to make prices cheaper or **make premium products more attractive** and exclusive, this will help **entice Customers to buy the premium**, upgraded, option, hence increasing transaction amount traffic on our platform.

We can achieve this by **partnering with popular brands** like Retrospot, Design, and Polkadot **to release special products exclusive to the brand** and **hype up the demand** for these premium options. Then, we can **include small discounts** for premium products to make Customers feel that they are getting a good deal if they buy the premium option – Customers go for the brand and feel happy when they get a discount for getting the upgraded and better version of the product.

I also uncovered that **large types of furniture are hard to sell off** as they are the worst-performing products generally. Hence, we could perhaps **include free delivery and installation for big, bulky, and expensive products** to convince customers to purchase by **ensuring comfort and ease of setup for customers**. Additionally, we can **compete with other furniture providers** by **offering competitive prices** on our expensive and premium options.

Cross-Selling:

The **top selling products are kitchenware and food-related products** like Cultery, Cake, and Lunch-related products.

There are also products like wraps and decorations, which are for **celebrating events** like birthday parties or festivals like Christmas. We can **bundle relevant products together to persuade customers to buy more**.

Bundling products benefits **Customers as they may forget to purchase an item when they have to individually find the products** (e.g. forgot to buy Ornaments for a Christmas tree). **Bundling relevant products helps prevent such hindsight**, while also **improving customer experience** as they don't have to manually find everything which is tedious, which boosts our sales when Customers buy a complete set.

Additionally, **based on data, it seems that cross-selling will be an effective method** to boost sales as the top performing product type is 'SET', meaning customers have a tendency of choosing a SET than individual packing.

Possible bundles:

1. **Products that work together**: Lunch Set (Lunch Box, Lunch Bag, Cutlery Set)
2. Set of **items with the same brand or design**, for collectors: Retrospot/Polkadot LunchBox, Bag, etc.
3. **Festival Pack**: Christmas Decorations + Ornaments + Christmas Tree + Wrapping Paper
4. **Bundle wrapping paper with purchases** that are **possibly Christmas gifts** during November and December
5. **Complimentary gifts** to entice customers to buy: Product + Hard-to-sell products like a Sketchbook

Recommendations to increase Sales in Spain

After understanding the Top Products:

1. Get more cutlery, lunch boxes and bags suppliers with **unique and interesting designs** to tempt Customers to buy the trendiest ones.
2. **Sell more types of Retrospot products** as this brand is in high demand.
3. Focus on **creating products with Polkadot design** as it's favored by Customers in Spain.

Create more designs and variations of the same product:

- **Give customers more choices to choose from**. Maybe they might even buy multiple because they can't decide which one they like more. Top Suppliers sell the same product with different designs and brands.

Tap into the Middle-Aged Adults market in Spain:

- Sell more **products targetted towards adults**
- Attract more adults to use our platform for purchasing goods

Platform performance and Customer Loyalty:

- Hire **developers to revamp the platform frequently** to ensure the design is modern and appealing.
- Hire **reliable engineers to maintain the servers and keep it running 24/7 without crashing**.
- Implement a **reward system to get customers to consistently use our platform** to buy items – discount coupons if Customers make frequent purchases.

Increase customer base:

- **Advertise** our platform in Spain to let people know about our existence.
- Run **campaigns to inculcate the habit of online shopping**, to create a lifestyle of online shopping for people of all ages.
- We should **target middle-aged adults** and **homeowners looking for furniture and appliances**.

Attract people who are looking for appliances and furniture to our platform:

- Since sales of large furniture are poor, we should advertise and convince people to use our platform to buy furniture for their new home or to replace their broken furniture.
- To attract them, we can either try cross-selling (bundling related furniture like tables and chairs), giving discounts and freebies, or giving free delivery and installation services to customers.

Since sales are low during weekends, **release discounts or bundle deals during weekends** to boost sales, and **consider opening the platform on Saturdays too**.

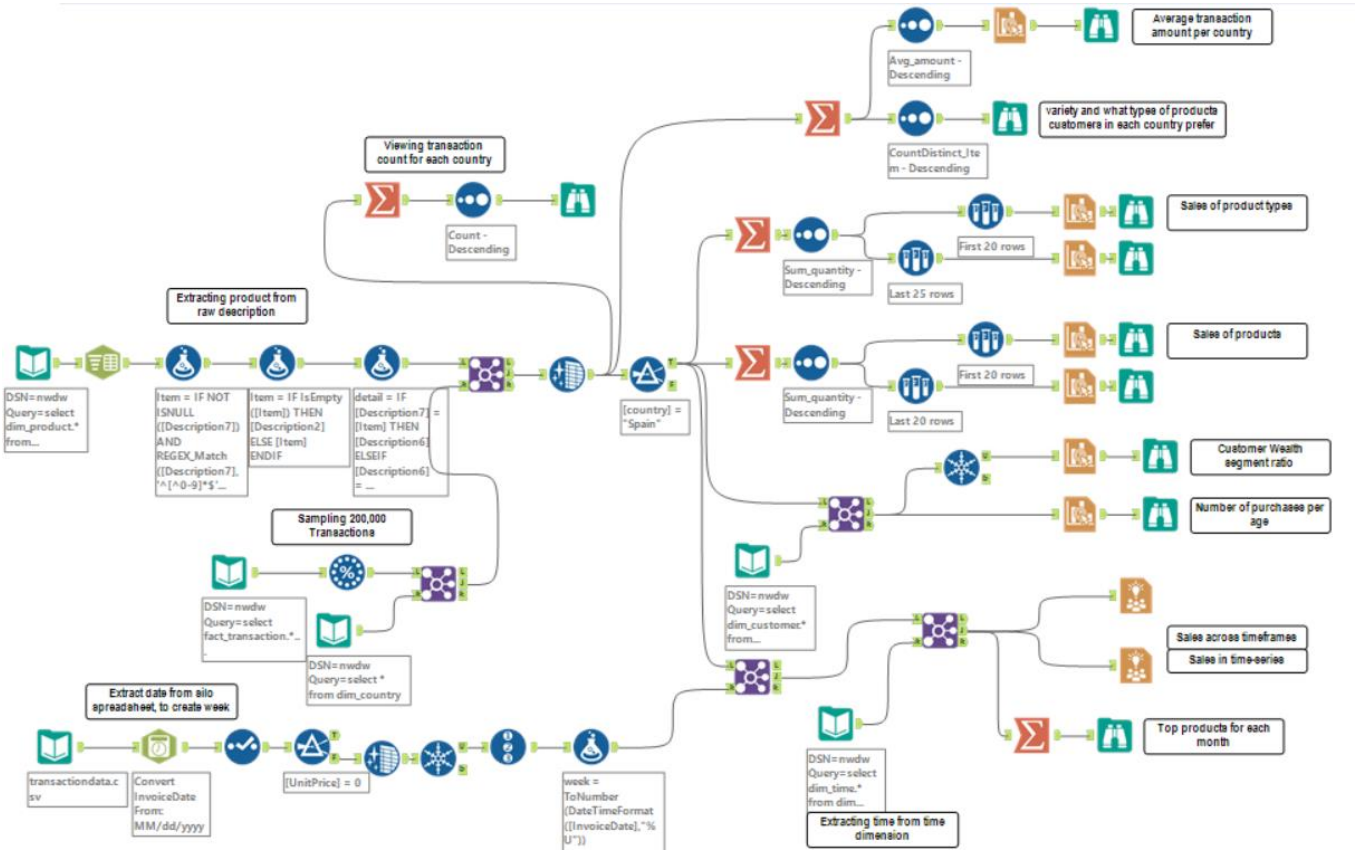
Lastly, we can try **implementing bundle deals** like Back-to-school Bundle (Lunch box, bag, and cutleries), Birthday Bundle (Cake products) and Christmas Set to convince Customers to use our platform to buy products for celebrating events and festivals since we have already helped them package all the neccessary items needed which **reduces their hassles** for preparing for these events.

Conclusion:

Sales performance in Spain **can be improved** as it is a relatively untapped market. Some recommendations I propose include **Advertising** to reach audiences and persuade them to try online shopping, giving **discounts** like price reductions or bundled products to entice Users to buy more things, and **hiring capable and reliable developers and engineers** to run and maintain the platform.

Additionally, we should work with Suppliers to **produce trending products that appeal to Adults**, and prioritize our focus on working with Top Suppliers and Suppliers of **kitchen products** and get them to **push out new products frequently** to **attract Customers to spend**. Lastly, we should **include free shipping and installation for big and bulky products.** By tapping into the untouched market of Middle-Aged Adults and Furniture and Appliances Customers, the sales of our retail platform can definitely be improved.

Workflow Overview:



End of Report