This support document file contains three main sections:

* Data cleaning process with detailed description (p.1).
* SQL queries used for the report (p.1).
* Other tables & figures that were not included in Main report (p.6).

**Data Cleaning**

For exploration step, the original table (lineitems\_sample.csv) were loaded into PostgreSQL server with the name ‘line\_items’ under ‘lixan23’ user, ‘nlab’ database and ‘asa’ schema. An ‘ALTER’ query (SQL Query 1) was applied to transform all spending fields from ‘money’ type to ‘numeric’.

To ensure the data quality, a sanity check was carried out. Overall, no null or replicate value found (SQL Query 2). However, there were 865 lines detected with negative value for ‘quantity’ and ‘spend’ (SQL Query 3 – Appendix).

Since there was no extra information on whether these lines were correctly recorded (for example, in refund case), and their contribution to the dataset was relatively small (the maximum loss observed in ‘LOTTERY’, with negative lines accounted for around 5% of total records), they were removed from the analysis base (SQL Query 5 – Appendix) to create a new table.

The detailed breakdown for impacted category can be found in SQL Query 4.

**SQL Queries**

|  |
| --- |
| **SQL Query 1.** Convert monetary value to numerical |
| **Formula:** Alter monetary value in ‘spend’ column to numeric |
| **SQL Query:**  ALTER TABLE asa.line\_items  ALTER COLUMN spend TYPE NUMERIC |
| **Result:** revised table |

|  |
| --- |
| **SQL Query 2.** Checking missing value |
| **Formula:** Find NULL value in all columns of original transaction data |
| **SQL Query:**  SELECT \*  FROM asa.line\_items  WHERE customer\_number IS NULL  OR purchase\_time IS NULL  OR product\_id IS NULL  OR category IS NULL  OR quantity IS NULL  OR spend IS NULL |
| **Result:** no NULL value found |

|  |
| --- |
| **SQL Query 3.** Checking negative value |
| **Formula:** Count lines with negative value in ‘quantity’ or ‘spending’ of original transaction data |
| **SQL Query:**  SELECT COUNT(\*)  FROM asa.line\_items  WHERE quantity<=0  OR spend<=0 |
| **Result:** 865 records |

|  |
| --- |
| **SQL Query 4.** Checking categories impacted by negative value |
| **Formula:** Calculate % of line with negative value in ‘quantity’ or ‘spending’  In each category |
| **SQL Query:**  WITH  missing AS (  SELECT category, COUNT(\*) AS miss  FROM asa.line\_items  WHERE quantity<=0 OR spend<=0  GROUP BY 1),  all\_ca AS (  SELECT category, COUNT(\*) AS all\_ca  FROM asa.line\_items  GROUP BY 1)  SELECT category, miss\*100.0/all\_ca AS missing\_per  FROM missing JOIN all\_ca USING (category) |
| **Result: (%)**  "BAKERY" 0.00087051900342984487  "CONFECTIONARY" 0.00116738558162068140  "DELI" 0.00324801870858776147  "DRINKS" 0.00266880170803309314  "GROCERY\_HEALTH\_PETS" 0.00243510446598159061  "LOTTERY" 5.0515958723302136  "MEAT" 0.00175904589350736161  "SOFT\_DRINKS" 0.02305475504322766571 |

|  |
| --- |
| **SQL Query 5.** Creating new, clean data |
| **Formula:** Create new table removing negative value |
| **SQL Query:**  CREATE TABLE asa.line\_items\_clean AS (  SELECT \*  FROM asa.line\_items  WHERE quantity>0  AND spend>0) |
| **Result:** ‘line\_items\_clean.csv’ |

|  |
| --- |
| **SQL Query 6.** Monthly tracking for total sales, customer base |
| **Formula:** Sum of spend and count of unique customers each operation month |
| **SQL Query:**  SELECT DATE\_TRUNC ('month',purchase\_time),  SUM(spend),  COUNT(DISTINCT customer\_number)  FROM asa.line\_items\_clean  GROUP BY 1 |
| **Result:**  "2007-03-01 00:00:00" 422516.28 2985  "2007-04-01 00:00:00" 385330.73 2948  "2007-05-01 00:00:00" 392005.01 2895  "2007-06-01 00:00:00" 380508.87 2847  "2007-07-01 00:00:00" 371211.94 2857  "2007-08-01 00:00:00" 361774.22 2802 |

|  |
| --- |
| **SQL Query 7.** Feature generation – Proportion of spending for categories |
| **Formula:**   * Compute total spending of each customer * Calculate % of cate spending over total spending of each customer |
| **SQL Query:**  WITH  total\_spend AS (  SELECT customer\_number, SUM(spend) AS total\_spend  FROM asa.line\_items\_clean  GROUP BY 1),  cate\_spend AS(  SELECT customer\_number, LOWER(category) AS category, SUM(spend) AS cate\_spend  FROM asa.line\_items\_clean  GROUP BY 1,2  ORDER BY 1,2)  SELECT a.customer\_number, category, cate\_spend\*100/total\_spend as cate\_contribution  FROM total\_spend a JOIN cate\_spend USING (customer\_number) |
| **Result:** ‘%cate\_clean\_uncross.csv’ |

|  |
| --- |
| **SQL Query 8.** Feature generation – Shopping behaviour features |
| **Formula:**  Create the data by:   * Compute total spending, total quantity, total visit, basket value, basket product (unique product per basket), unit cost, weekly spending, weekly visit, days from last purchase, number of categories, number of products, number of products per category * Compute spending change:  1. Compute spending in first 3 month 2. Compute spending in last 3 month 3. Take the division of 2/1  * Compute % weekend visit:  1. Compute the day of all transaction 2. Compute the number of weekday and weekend visits for each customer 3. Create table for weekday and weekend visit separately 4. Full outer join 2 tables to make sure no missing record for customers solely purchase on weekend or weekday 5. Take the division of weekend visits \*100/ (weekend + weekday visits) |
| **SQL Query:**  WITH  rfm AS(  SELECT customer\_number,  SUM (spend) AS total\_spend,  SUM(quantity) AS total\_quantity,  COUNT(DISTINCT purchase\_time) AS total\_visit,  SUM (spend)/ COUNT(DISTINCT purchase\_time) AS basket\_value,  COUNT(DISTINCT product\_id)\*1.0/COUNT(DISTINCT purchase\_time) AS basket\_product,  SUM (spend)/ SUM(quantity) AS unit\_cost,  SUM(spend)/(EXTRACT(week FROM MAX(purchase\_time))- EXTRACT(week from MIN(purchase\_time))+1) AS weekly\_spend,  COUNT(DISTINCT purchase\_time)/(EXTRACT(week FROM MAX(purchase\_time))- EXTRACT(week from MIN(purchase\_time))+1) AS weekly\_visit,  EXTRACT (DAY FROM('2007-08-31'-MAX(purchase\_time))) AS last\_purchase,  COUNT(DISTINCT category) AS num\_cate,  COUNT(DISTINCT product\_id) AS num\_pro,  COUNT(DISTINCT product\_id)/COUNT(DISTINCT category) AS pro\_per\_cate  FROM asa.line\_items\_clean  GROUP BY 1),    spend\_change AS(  WITH  first\_half AS (  SELECT customer\_number, SUM(spend) AS f3m  FROM asa.line\_items\_clean  WHERE EXTRACT(MONTH from purchase\_time) in (3,4,5)  GROUP BY 1),  sec\_half AS (  SELECT customer\_number, SUM(spend) AS l3m  FROM asa.line\_items\_clean  WHERE EXTRACT(MONTH from purchase\_time) in (6,7,8)  GROUP BY 1)  SELECT first\_half.customer\_number, l3m\*1.0/f3m AS spend\_change  FROM first\_half FULL OUTER JOIN sec\_half USING (customer\_number)),  visit\_time AS(  WITH  date\_time AS (  SELECT customer\_number,  EXTRACT (isodow from purchase\_time) AS day  FROM asa.line\_items),  visit\_frequency AS (  SELECT customer\_number,  CASE WHEN (day>=5)  THEN 'weekend'  ELSE 'weekday'  END AS visit, COUNT(\*) AS frequency  FROM date\_time  GROUP BY 1,2),  weekend AS (  SELECT customer\_number, visit, frequency  FROM visit\_frequency  WHERE visit='weekend'),  weekday AS (  SELECT customer\_number, visit, frequency  FROM visit\_frequency  WHERE visit='weekday'),  join\_table AS (  SELECT a.customer\_number, b.customer\_number AS cross\_check,  CASE WHEN a.frequency IS null  THEN 0  ELSE a.frequency END AS weekend,  CASE WHEN b.frequency IS null  THEN 0  ELSE b.frequency END AS weekday  FROM weekend a FULL OUTER JOIN weekday b USING (customer\_number))  SELECT CASE WHEN customer\_number IS null  THEN cross\_check ELSE customer\_number  END AS customer\_number,  weekend\*100.0/(weekend+weekday) AS weekend\_percent  FROM join\_table)    SELECT \*  FROM rfm JOIN spend\_change USING (customer\_number)  JOIN visit\_time USING (customer\_number) |
| **Result:** ‘all\_except\_cate\_final.csv’ |

|  |
| --- |
| **SQL Query 9.** Monthly tracking for total sales by cluster |
| **Formula:**   * Upload cluster classification file ‘cluster\_id’ to SQL server under ‘asa.clusters’ * Compute the sum of spend group by month and customer number |
| **SQL Query:**  SELECT cluster, EXTRACT (month from purchase\_time), SUM(spend)  FROM asa.line\_items\_clean  JOIN asa.clusters USING(customer\_number)  GROUP BY 1,2 |
| **Result:** Figure 12 Final Report |

**Other tables & figures**

**Table 13.** Generated Features

Table

Description automatically generated

**Figure 14.** Features’ Correlation Matrix

Chart, scatter chart

Description automatically generated

**Table 15.** Selected features

Table

Description automatically generated

**Figure 16.** Features’ histogram before (grey) and after (blue) Power Transformation & Standardization

**A picture containing text, crossword puzzle

Description automatically generated**

**Figure 17.** Silhouette Visualization

Chart, funnel chart

Description automatically generatedChart, funnel chart

Description automatically generated

Chart

Description automatically generated with medium confidenceChart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidenceChart

Description automatically generated

**Table 18.** Clusters’ average (mean) value for all features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Features** | **Super Customers** | **Bustling Homemakers** | **One-stop shoppers** | **Economical Neighbours** | **Potential Lapsers** |
| **BEHAVIOURS** | | | | | |
| total\_spend | 1403.78 | 761.31 | 712.51 | 576.67 | 263.45 |
| total\_quantity | 1000.58 | 631.85 | 370.74 | 590.90 | 205.37 |
| total\_visit | 115.69 | 36.17 | 66.43 | 77.85 | 23.06 |
| basket\_size | 13.78 | 23.73 | 12.29 | 8.20 | 14.48 |
| basket\_product | 3.49 | 8.50 | 2.42 | 2.94 | 6.42 |
| unit\_cost | 1.44 | 1.22 | 2.03 | 1.00 | 1.34 |
| weekly\_spend | 52.43 | 29.59 | 27.05 | 21.66 | 14.25 |
| weekly\_visit | 4.32 | 1.40 | 2.52 | 2.93 | 1.19 |
| last\_purchase | 1.07 | 4.81 | 3.03 | 1.27 | 35.03 |
| num\_cate | 17.25 | 14.89 | 14.87 | 14.65 | 13.82 |
| num\_pro | 337.26 | 269.43 | 131.82 | 202.45 | 110.67 |
| pro\_per\_cate | 19.11 | 17.56 | 8.34 | 13.30 | 7.40 |
| spend\_change | 1.03 | 1.11 | 0.98 | 1.00 | 0.45 |
| weekend\_percent | 41.07 | 49.67 | 39.60 | 37.99 | 40.90 |
| **% CATEGORIES SPENDING** | | | | | |
| bakery | 4.57 | 5.43 | 3.77 | 7.38 | 5.38 |
| cashpoint | 7.33 | 1.29 | 12.64 | 2.03 | 4.11 |
| confectionary | 6.79 | 9.31 | 4.27 | 9.84 | 9.42 |
| dairy | 8.32 | 11.53 | 6.50 | 12.73 | 10.23 |
| deli | 1.65 | 1.79 | 0.97 | 3.19 | 2.33 |
| discount\_bakery | 0.23 | 0.22 | 0.08 | 0.46 | 0.49 |
| drinks | 9.15 | 5.16 | 14.70 | 3.15 | 8.34 |
| frozen | 4.15 | 6.07 | 2.68 | 5.47 | 5.02 |
| fruit\_veg | 7.61 | 12.70 | 5.07 | 12.09 | 9.43 |
| grocery\_food | 6.72 | 10.69 | 4.50 | 9.85 | 8.44 |
| grocery\_health\_pets | 7.41 | 10.60 | 4.37 | 8.08 | 8.67 |
| lottery | 3.01 | 0.41 | 3.34 | 1.34 | 1.36 |
| meat | 6.68 | 9.69 | 4.70 | 6.68 | 7.52 |
| newspapers\_magazines | 2.45 | 1.49 | 1.58 | 3.93 | 2.26 |
| practical\_items | 0.26 | 0.34 | 0.18 | 0.26 | 0.39 |
| prepared\_meals | 4.29 | 6.12 | 3.01 | 5.05 | 5.26 |
| seasonal\_gifting | 0.82 | 0.97 | 0.49 | 0.84 | 1.49 |
| soft\_drinks | 3.27 | 3.06 | 1.56 | 3.51 | 3.12 |
| tobacco | 14.29 | 1.90 | 24.89 | 2.20 | 5.21 |
| world\_foods | 1.00 | 1.24 | 0.70 | 1.93 | 1.54 |