```
/*
 Task 1 of the Quantium Virtual Internship.
 Goal: Clean Data and Perform Analysis on customer segments and their chip
 purchasing behaviour
 I have chosen to perform this task in SQL because I'm not quite proficient in Python or
 R.
 Skills used: Converting Data Types, Manipulation syntax, String Expressions, Windows
 Functions, Joins, CTE
*/

/* DATA VALIDATION AND DATA CLEANING */
-- Convert DATE column to a date format, CSV and Excel integer dates begin on 30 Dec
1899
ALTER TABLE SalesPerformance.transaction_data
ADD COLUMN DATE_Converted date
   GENERATED ALWAYS AS (DATE_ADD('1899-12-30', INTERVAL DATE DAY))
   STORED;

-- examine the PROD_NAME column, remove digit and special characters, sort by
frequency of occurrence
-- remove size, '[\/&]| +': matches '/','/' or '&' or multiple spaces
SELECT DISTINCT
TRIM(REPLACE(REGEXP_REPLACE(REGEXP_REPLACE(PROD_NAME,
'[0-9]+g',''),'[\/&]+| +',' '),'   ',' ')) as filter_PROD, COUNT(*) as frequency
FROM SalesPerformance.transaction_data
GROUP BY 1
ORDER BY 2 DESC;

-- most frequency: Cobs Popd SwtChlli SrCream Chips and Cobs Popd Sea Salt Chips
-- add a new column with cleaned PROD_NAME

ALTER TABLE SalesPerformance.transaction_data
ADD COLUMN Cleaned_PROD TEXT
   GENERATED ALWAYS AS
   (TRIM(REPLACE(REGEXP_REPLACE(REGEXP_REPLACE(PROD_NAME,
   '[0-9]+g',''),'[\/&]+| +',' '),'   ',' '))) STORED;

-- check if products are in chips category and filter out salsa, check null values and
```

```
possible outliers
SELECT *
FROM SalesPerformance.transaction_data
WHERE PROD_NAME NOT LIKE '%chip%';

DELETE FROM SalesPerformance.transaction_data
WHERE STORE_NBR IS NULL OR
   LYLTY_CARD_NBR IS NULL OR
   TXN_ID IS NULL OR
   PROD_NBR IS NULL OR
   Cleaned_PROD IS NULL OR
   PROD_QTY IS NULL OR
   TOT_SALES IS NULL;

-- there's no null values in all columns
-- remove salsa products
DELETE FROM SalesPerformance.transaction_data
WHERE PROD_NAME LIKE '%salsa%';

-- check statistics such as mean, min and max values for each feature to detect outliers
SELECT MIN(PROD_QTY), MAX(PROD_QTY), AVG(PROD_QTY), MIN(TOT_SALES),
MAX(TOT_SALES), AVG(TOT_SALES)
FROM SalesPerformance.transaction_data;

-- Max values are significantly far from the Avg, could have outliers
-- filter dataset to find outlier
SELECT *
FROM SalesPerformance.transaction_data
WHERE PROD_QTY >= 100 AND PROD_QTY <= 200;

-- There are two transactions where 200 packets of chips are bought and both came
from the same customer: Card_nbr 226000
-- Let's see if the customer has had other transactions
SELECT *
FROM SalesPerformance.transaction_data
WHERE LYLTY_CARD_NBR = 226000;

-- looks like this cus is not a retail cus, might be buying chips for commercial purposes
```

```sql
instead -> remove this customer in dataset
DELETE
FROM SalesPerformance.transaction_data
WHERE LYLTY_CARD_NBR = 226000;

-- check number of transaction over time to see if there's missing data
SELECT DATE_Converted, COUNT(*) AS num_of_transaction
FROM SalesPerformance.transaction_data
GROUP BY DATE_Converted
ORDER BY DATE_Converted;

-- 364 rows return -> 1 missing date.
WITH t AS
    (SELECT DATE_Converted, COUNT(*) AS num_of_transaction,
    (DATEDIFF(DATE_Converted, LAG(DATE_Converted) OVER (ORDER BY
    DATE_Converted))) as day
    FROM SalesPerformance.transaction_data
    GROUP BY DATE_Converted
    ORDER BY DATE_Converted)

SELECT * FROM t
WHERE day > 1
ORDER BY DATE_Converted;

-- 2018-12-25 is the missing date -> This is due to shops being closed on Christmas day
-- create pack size from PROD_NAME. The largest size is 380g and the smallest size is
70g
ALTER TABLE SalesPerformance.transaction_data
ADD COLUMN PACK_SIZE INT
    GENERATED ALWAYS AS (REGEXP_SUBSTR(PROD_NAME, '[0-9]+(?=g)'))
    STORED;

-- create brand name: use the first word in PROD_NAME
ALTER TABLE SalesPerformance.transaction_data
ADD COLUMN BRAND TEXT
    GENERATED ALWAYS AS (SUBSTRING_INDEX(PROD_NAME, ' ',1)) STORED;

-- clean brand name, fix the spelling mistake
```

```
UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'Dorito','Doritos')
WHERE PROD_NAME LIKE 'Dorito%'
  AND PROD_NAME NOT LIKE 'Doritos%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'Infzns','Infuzions')
WHERE PROD_NAME LIKE 'Infzns%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'GrnWves','GrainWaves')
WHERE PROD_NAME LIKE 'GrnWves%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'Grain Waves','GrainWaves')
WHERE PROD_NAME LIKE 'Grain Waves%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'RRD','Red')
WHERE PROD_NAME LIKE 'RRD%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'Smith','Smiths')
WHERE PROD_NAME LIKE 'Smith%'
AND PROD_NAME NOT LIKE 'Smiths%';

UPDATE SalesPerformance.transaction_data
SET PROD_NAME = REPLACE(PROD_NAME,'Snbts','Sunbites')
WHERE PROD_NAME LIKE 'Snbts%';

-- examine distributions of key columns
SELECT BRAND, COUNT(*) as count, ROUND(100 * COUNT(*) / SUM(COUNT(*))
OVER (), 2) AS percent
FROM SalesPerformance.transaction_data
GROUP BY BRAND ORDER BY percent DESC;

SELECT
  CASE
```

```sql
        WHEN TOT_SALES <= 5 THEN '0-5'
        WHEN TOT_SALES > 5 AND TOT_SALES <= 10 THEN '5-10'
        WHEN TOT_SALES > 10 AND TOT_SALES <= 15 THEN '10-15'
        WHEN TOT_SALES > 15 AND TOT_SALES <= 20 THEN '15-20'
        WHEN TOT_SALES > 20 AND TOT_SALES <= 25 THEN '20-25'
        WHEN TOT_SALES > 25 AND TOT_SALES <= 30 THEN '25-30'
    END AS Sales_Range,
    COUNT(*) as count
FROM SalesPerformance.transaction_data
GROUP BY 1
ORDER BY MIN(TOT_SALES);

-- Merge transaction data to customer data.
CREATE TABLE SalesPerformance.cleaned_SalesData AS
(SELECT DATE_Converted as DATE,
        STORE_NBR,
        t.LYLTY_CARD_NBR,
        LIFESTAGE,
        PREMIUM_CUSTOMER,
        TXN_ID,
        PROD_NBR,
        Cleaned_PROD AS PRODUCT,
        BRAND,
        PACK_SIZE,
        PROD_QTY AS QUANTITY,
        TOT_SALES
    FROM SalesPerformance.transaction_data t
        JOIN SalesPerformance.purchase_behaviour p
        ON t.LYLTY_CARD_NBR = p.LYLTY_CARD_NBR
    ORDER BY 1);

-- summary statistic
WITH t AS
(SELECT PREMIUM_CUSTOMER,
        TOT_SALES / QUANTITY AS price_per_unit
FROM SalesPerformance.cleaned_SalesData
WHERE QUANTITY > 0
 )
```

```
SELECT
  PREMIUM_CUSTOMER,
  AVG(price_per_unit) AS Avg_price,
  STDDEV_SAMP(price_per_unit) AS sd_price,
  COUNT(*) AS num_of_customer
FROM t
GROUP BY PREMIUM_CUSTOMER;

-- check for missing customer details
SELECT *
FROM SalesPerformance.cleaned_SalesData
WHERE LIFESTAGE IS NULL OR PREMIUM_CUSTOMER IS NULL;

-- no null values, so all our customers in the transaction data has been accounted for in
the customer dataset.

/* DATA ANALYSIS ON CUSTOMER SEGMENTS */
-- How many customer are in each segment?
SELECT PREMIUM_CUSTOMER,
    LIFESTAGE,
    COUNT(*) AS num_of_transaction,
    COUNT(DISTINCT LYLTY_CARD_NBR) AS num_of_customer,
    ROUND(100 * COUNT(DISTINCT LYLTY_CARD_NBR) / SUM(COUNT(DISTINCT
    LYLTY_CARD_NBR)) OVER (), 2) AS pct_of_customer_base
FROM SalesPerformance.cleaned_SalesData
GROUP BY 1,2
ORDER BY pct_of_customer_base DESC;

-- 40.31% Mainstream (also accounted for highest number of purchase), 33.68% Budget
and 26.02% Premium
-- Mainstream - Young Singles/Couples accounted for the highest proportion of total
customer base (~11.11%)

-- calculate average chip price by product
WITH t AS
    (SELECT PRODUCT, ROUND(SUM(TOT_SALES)/SUM(QUANTITY),2) AS PRICE
    FROM SalesPerformance.cleaned_SalesData
    GROUP BY PRODUCT)
```

```sql
SELECT MIN(PRICE), MAX(PRICE), ROUND(AVG(PRICE),2)
FROM t;

-- the cheapest is $1.7, the highest price is $6.39 and avg.price is $3.52

-- Who spends the most on chips (total sales)?
-- calculate total sales, number of customer and average buying units by LIFESTAGE
and PREMIUM_CUSTOMER
SELECT PREMIUM_CUSTOMER,
    LIFESTAGE,
    ROUND(SUM(TOT_SALES),2) as Total_Sales,
    COUNT(DISTINCT LYLTY_CARD_NBR) AS Num_of_customer,
    AVG(QUANTITY) AS Avg_Unit,
    ROUND(SUM(TOT_SALES)/SUM(QUANTITY),2) AS Price_Per_Unit,
    ROUND(100 * SUM(TOT_SALES) / SUM(SUM(TOT_SALES)) OVER (), 2) AS
    Pct_of_Total_Sales
FROM SalesPerformance.cleaned_SalesData
GROUP BY PREMIUM_CUSTOMER, LIFESTAGE
ORDER BY Pct_of_Total_Sales DESC, Num_of_customer DESC;

-- Sales come mainly from Budget - Older Families, Mainstream - young singles/couples,
and Mainstream - retirees
-- Older families and young families in general buy more chips per customer
-- more customers buying chips are Mainstream - young singles/couples and
Mainstream - retirees. so these 2 segments contribute more in total sales, but this is not
a major driver for the highest sales in Budget - Older families segment

-- calculate average price per unit chips bought by customer segment
SELECT PREMIUM_CUSTOMER,
    LIFESTAGE,
    SUM(QUANTITY) AS Total_Quantity,
    SUM(TOT_SALES) AS Total_Sales,
    ROUND(SUM(TOT_SALES)/SUM(QUANTITY),2) AS Price_Per_Unit,
    AVG(QUANTITY) AS Avg_Unit
FROM SalesPerformance.cleaned_SalesData
GROUP BY 1,2
ORDER BY Price_Per_Unit DESC, Avg_Unit DESC;
```

```
-- ~1.8-1.9 pack of chips are bought per each segment. the difference in average chips
units being bought isn't significant
-- Mainstream mid-age and young singles/couples are more willing to pay more per
packet of chips

-- calculate average spending per customer by segment
SELECT PREMIUM_CUSTOMER,
       LIFESTAGE,
       COUNT(DISTINCT LYLTY_CARD_NBR) AS Customer_base,
       ROUND(SUM(TOT_SALES),2) AS Total_Sales,
       ROUND(SUM(TOT_SALES)/COUNT(DISTINCT LYLTY_CARD_NBR),2) AS
       Avg_Spend_Per_Customer
FROM SalesPerformance.cleaned_SalesData
GROUP BY PREMIUM_CUSTOMER, LIFESTAGE
ORDER BY Avg_Spend_Per_Customer DESC, Customer_base DESC;

-- Customers in Older Families - both Mainstream and Budget segments have highest
average spend for chips

-- Let's deep dive into target 2 customer segments that contribute the most to sales:
Budget - Older Families and Mainstream - young singles/couples
-- For instance, let's find out if they tend to buy a particular brand of chips.
SELECT BRAND,
   SUM(QUANTITY) AS Total_qty,
   SUM(TOT_SALES) AS Total_Sales
FROM SalesPerformance.cleaned_SalesData
WHERE PREMIUM_CUSTOMER = 'Mainstream'
   AND LIFESTAGE LIKE 'YOUNG SINGLES%'
GROUP BY BRAND
ORDER BY 3 DESC;

-- Kettle, Doritos and Pringles are top3 most favorite chip brands among young
mainstream customers

SELECT BRAND,
       SUM(QUANTITY) AS Total_qty,
       SUM(TOT_SALES) AS Total_Sales
FROM SalesPerformance.cleaned_SalesData
```

```sql
WHERE PREMIUM_CUSTOMER = 'Budget'
   AND LIFESTAGE LIKE 'Older%Families%'
GROUP BY BRAND
ORDER BY 3 DESC;

-- For Budget - Older Families, they are Kettle, Smiths and Pringles
-- brand Kettle was the first choice of both young singles/couples and older families

-- find out if our target segment tends to buy larger packs of chips
SELECT PREMIUM_CUSTOMER,
      LIFESTAGE,
      PACK_SIZE,
      COUNT(*) AS Total_Trans,
      SUM(TOT_SALES) AS Total_Sales
FROM SalesPerformance.cleaned_SalesData
WHERE PREMIUM_CUSTOMER = 'Mainstream'
   AND LIFESTAGE LIKE 'YOUNG SINGLES%'
GROUP BY 1,2,3
ORDER BY 1, 5 DESC, 4 DESC;

SELECT PREMIUM_CUSTOMER,
      LIFESTAGE,
      PACK_SIZE,
      COUNT(*) AS Total_Trans,
      SUM(TOT_SALES) AS Total_Sales
FROM SalesPerformance.cleaned_SalesData
WHERE PREMIUM_CUSTOMER = 'Budget'
   AND LIFESTAGE LIKE 'Older%Families%'
GROUP BY 1,2,3
ORDER BY 1, 5 DESC, 4 DESC;

-- Both Mainstream - young singles/couples and Budget - older families tend to buy medium-sized pack of chips (175g and 150g)

-- examine how well chips brands and store perform in sales
SELECT BRAND,
      SUM(QUANTITY) AS Total_Qty,
      ROUND(SUM(TOT_SALES),2) AS Total_Sales
```

```sql
FROM SalesPerformance.cleaned_SalesData
GROUP BY BRAND
ORDER BY 3 DESC;

SELECT STORE_NBR,
    COUNT(DISTINCT PRODUCT) AS Total_products,
    ROUND(SUM(TOT_SALES),2) AS Total_Sales,
    ROUND(100 * SUM(TOT_SALES) / SUM(SUM(TOT_SALES)) OVER (), 2) AS
    pct_of_total_sales
FROM SalesPerformance.cleaned_SalesData
GROUP BY STORE_NBR
ORDER BY 3 DESC;

-- Kettle, Doritos and Smiths are top 3 best-seller brands
-- We have 271 stores in total, Store 226 gains the best sales performance. Some stores
have little to no contribution to the total sales due to displaying very few products

SELECT *
FROM SalesPerformance.cleaned_SalesData;
-- Data exploration is now complete!

/* INSIGHT
  1. As we can see from customer behavior patterns:
    - Sales are coming mainly from budget-large households and younger mainstream
    Singles/Couples
    - The chips category is a mainstream (mass-market) category, not driven by
    premium buyers

  2. Budget–Older Families are the biggest sales drivers
    Although they are not the largest customer group, they still contribute the highest
    share of total chips sales with the largest product quantity (in total).
    This may suggest:
    - Their behaviour is volume-driven rather than premium-driven: they don’t pay the
    highest price per pack, but they buy more.
    - They buy in larger quantities per visit, and buy for households with multiple
    members.

  2. Mainstream customers dominate the shopper base (>40%), Young Singles/Couples
  and Retirees within this segment keep their total sales contribution high.

  3. Mainstream Singles/Couples are more willing to pay for chips
    Both young and mid-age Mainstream Singles/Couples show higher willingness to
    pay per packet compared to Budget or Premium customers in the same lifestage.
    - This may suggest they value brand, quality, or preferred flavors over price
    sensitivity.
    - There being fewer Premium mid-age and young singles/couples buying chips
    compared to their mainstream counterparts, despite being categorized as those
    buying higher-price products.
    -> maybe due to premium shoppers being more likely to buy healthy snacks, and
    when they buy chips, this is mainly for entertainment purposes rather than their own
    consumption.
    -> in the chips category, mainstream singles/couples show the real premium-like
    behaviour.

  5. For 2 customer segments that contribute the most to sales: Kettle & Pringles was
  their top choices, and they prefer medium-size packs of chips (175g and 150g)

*/
/* RECOMMENDATION
  To further increase sales, we should:
  - For Mainstream - Young singles/couples: Focus on increasing basket size, like
  flavor-driven bundles, new flavors innovation or brand loyalty rewards
  - For Budget - Older Families: Focus on value-driven promos, upselling and cross-
  selling
  - Consider promoting/prioritizing for placement the medium-sized packs of chips (175g
  and 150g)

*/
```