CUSTOMER SEGMENTATION USING RFM ANALYIS FOR A CHAIN OF RETAIL STORES

Problem Statement :

1. A chain of retail stores wants to launch a marketing campaign. Because of limited resources, they cannot target individual customers. This means that there is a need to optimize the targeting criteria so as to spend more on customers that will generate more revenue for the store.
2. Analyze the sales datas and arrive at actionable insights for growing the business

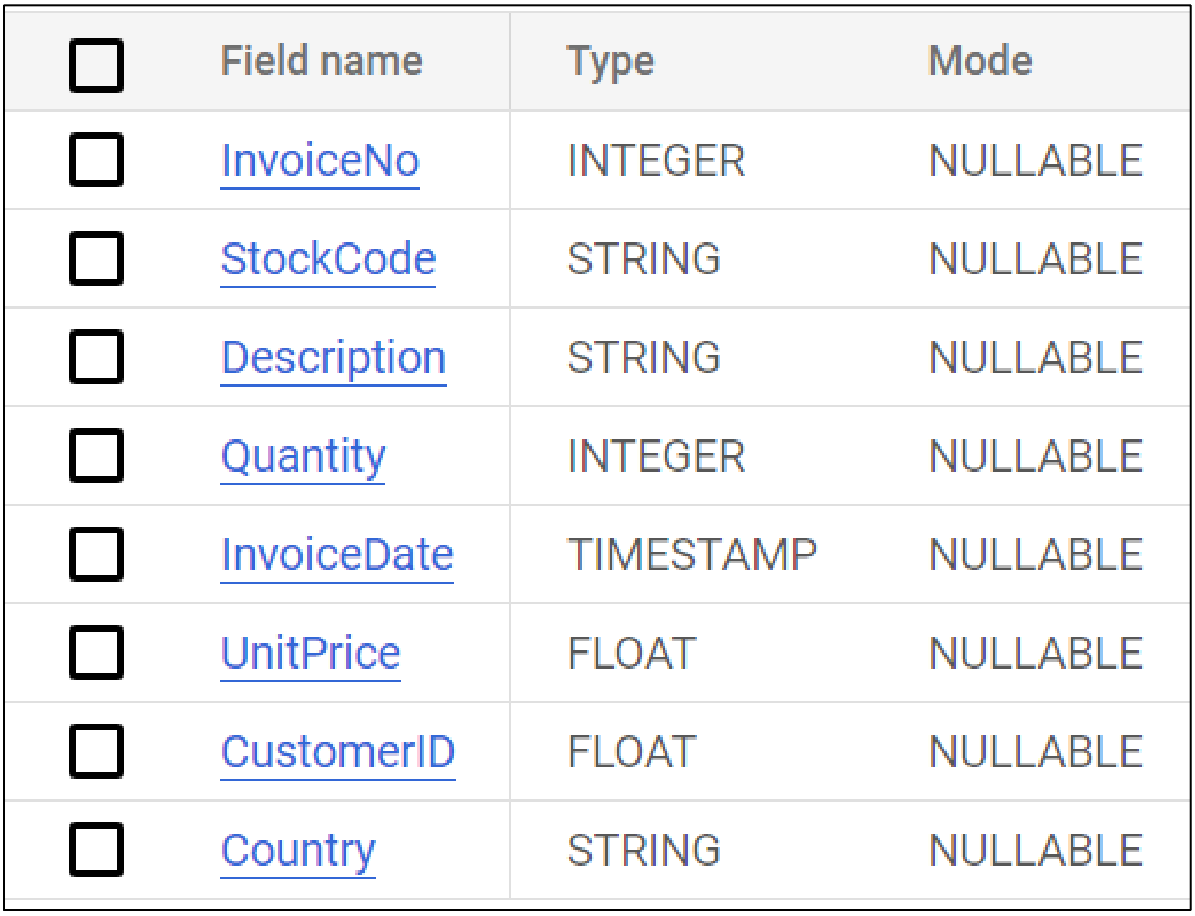
Analysis Plan :

1. Exploratory Data Analysis

2. RFM analysis for customer segmentation

Dataset :

Schema



□	Field name	Type	Mode
□	InvoiceNo	INTEGER	NULLABLE
□	StockCode	STRING	NULLABLE
□	Description	STRING	NULLABLE
□	Quantity	INTEGER	NULLABLE
□	InvoiceDate	TIMESTAMP	NULLABLE
□	UnitPrice	FLOAT	NULLABLE
□	CustomerID	FLOAT	NULLABLE
□	Country	STRING	NULLABLE

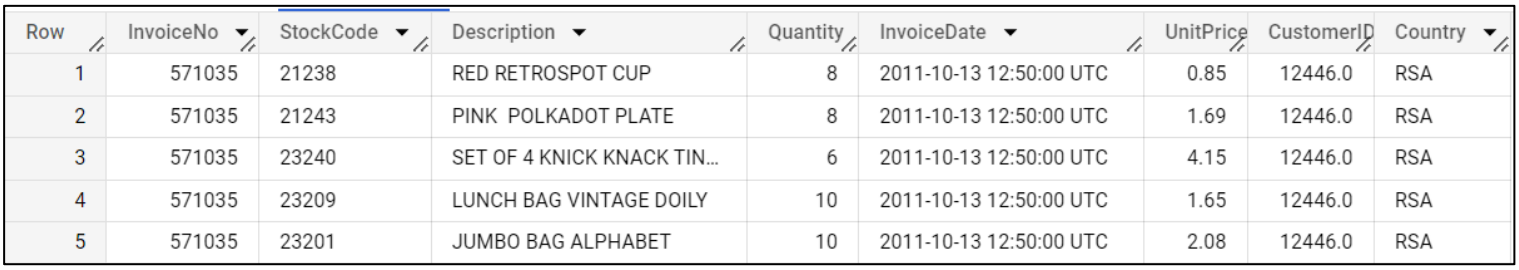
EXPLORATORY DATA ANALYSIS

Objective : View first 5 rows of the table

QUERY 1

```
SELECT
  *
FROM
  `retail.sales`
LIMIT 5
```

OUTPUT 1



Row	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	571035	21238	RED RETROSPOT CUP	8	2011-10-13 12:50:00 UTC	0.85	12446.0	RSA
2	571035	21243	PINK POLKADOT PLATE	8	2011-10-13 12:50:00 UTC	1.69	12446.0	RSA
3	571035	23240	SET OF 4 KNICK KNACK TIN...	6	2011-10-13 12:50:00 UTC	4.15	12446.0	RSA
4	571035	23209	LUNCH BAG VINTAGE DOILY	10	2011-10-13 12:50:00 UTC	1.65	12446.0	RSA
5	571035	23201	JUMBO BAG ALPHABET	10	2011-10-13 12:50:00 UTC	2.08	12446.0	RSA

Attribute Information :

•**InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

•**StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

•**Description:** Product (item) name. Nominal.

•**Quantity:** The quantities of each product (item) per transaction. Numeric.

•**InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.

•**UnitPrice:** Unit price. Numeric, Product price per unit in sterling.

•**CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

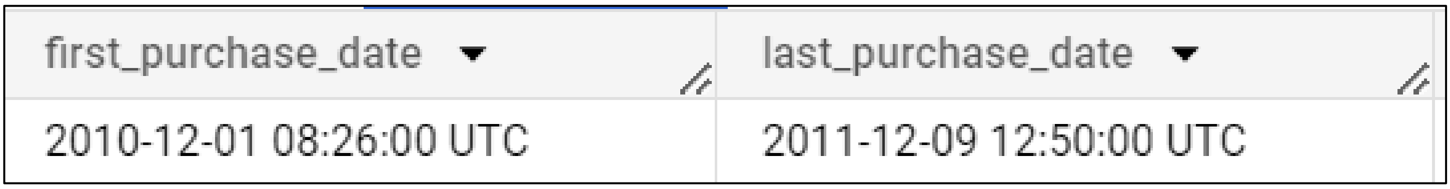
EXPLORATORY DATA ANALYSIS

Objective : Time period of the dataset

QUERY 2

```
SELECT
  MIN(InvoiceDate) as first_purchase_date,
  MAX(InvoiceDate) as last_purchase_date
FROM `retail.sales`
```

OUTPUT 2



first_purchase_date ▾	last_purchase_date ▾
2010-12-01 08:26:00 UTC	2011-12-09 12:50:00 UTC

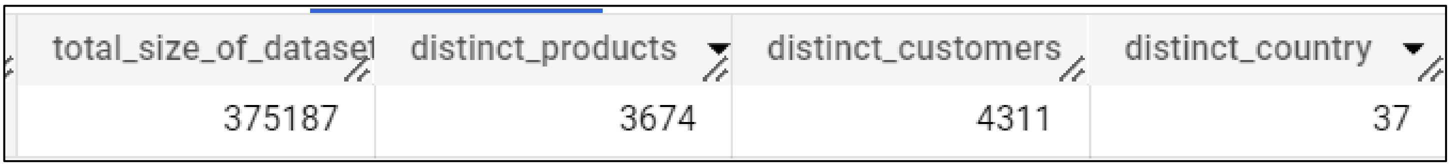
It is a 1 year dataset from December 1 2010 to December 9 2011

Objective : Size of dataset, count of distinct products, customers, country

QUERY 3

```
SELECT
  COUNT(*) as total_size_of_dataset,
  COUNT(DISTINCT Description) as distinct_products,
  COUNT(DISTINCT CustomerID) as distinct_customers,
  COUNT(DISTINCT Country) as distinct_country
FROM `retail.sales`
```

OUTPUT 3



total_size_of_dataset	distinct_products ▾	distinct_customers	distinct_country ▾
375187	3674	4311	37

There are 375187 records in the dataset, 3674 distinct products in the retail store, 4311 distinct customer records and datas from 37 different countries

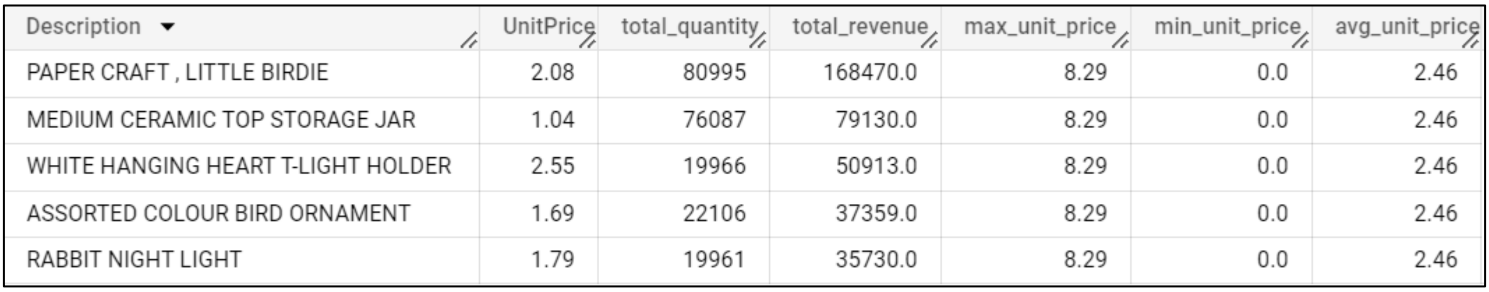
EXPLORATORY DATA ANALYSIS

Objective : Top 5 selling products in the retail stores

QUERY 4

```
SELECT
  Description,
  UnitPrice,
  SUM(Quantity) as total_quantity,
  ROUND(SUM(Quantity*UnitPrice),0) as total_revenue,
  ROUND(MAX(UnitPrice) OVER(),2) as max_unit_price,
  ROUND(MIN(UnitPrice) OVER(),2) as min_unit_price,
  ROUND(AVG(UnitPrice) OVER(),2) as avg_unit_price
FROM `retail.sales`
GROUP BY Description,UnitPrice
```

OUTPUT 4



Description	UnitPrice	total_quantity	total_revenue	max_unit_price	min_unit_price	avg_unit_price
PAPER CRAFT , LITTLE BIRDIE	2.08	80995	168470.0	8.29	0.0	2.46
MEDIUM CERAMIC TOP STORAGE JAR	1.04	76087	79130.0	8.29	0.0	2.46
WHITE HANGING HEART T-LIGHT HOLDER	2.55	19966	50913.0	8.29	0.0	2.46
ASSORTED COLOUR BIRD ORNAMENT	1.69	22106	37359.0	8.29	0.0	2.46
RABBIT NIGHT LIGHT	1.79	19961	35730.0	8.29	0.0	2.46

The unit price of top 5 revenue contributing products are less than or equal to the average unit price considering all the products in the retail store

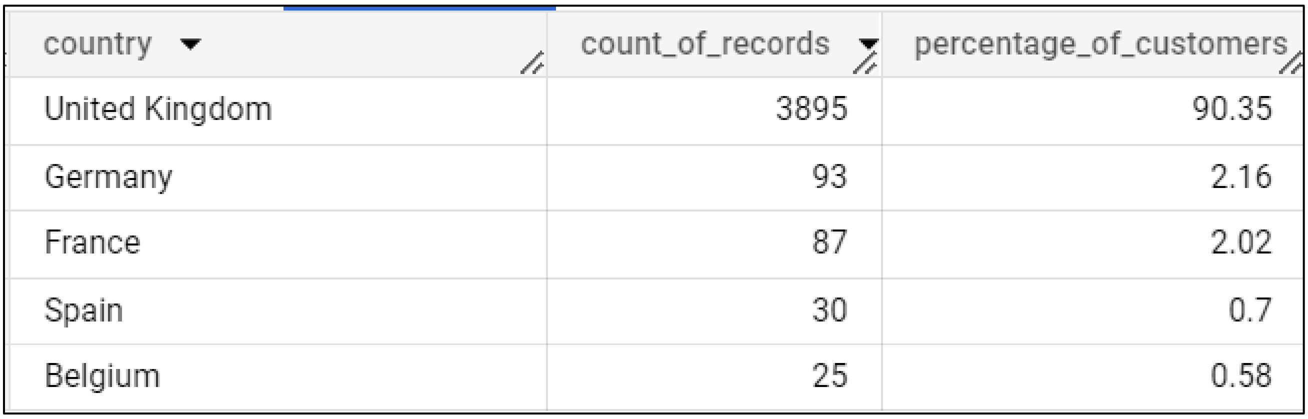
EXPLORATORY DATA ANALYSIS

Objective : Number of customer country wise

QUERY 5

```
WITH cte AS
  (SELECT
    DISTINCT Country,
    COUNT(DISTINCT CustomerID) as count_of_records
  FROM `retail.sales`
  GROUP BY Country),
  cte1 AS
  (SELECT
    COUNT(DISTINCT CustomerID) as total_count_of_records
  FROM `retail.sales`
  )
SELECT
  c.country,
  c.count_of_records ,
  ROUND(c.count_of_records*100/c1.total_count_of_records
,2) AS percentage_of_customers
FROM cte c, cte1 c1
ORDER BY c.count_of_records DESC
LIMIT 5
```

OUTPUT 5



country	count_of_records	percentage_of_customers
United Kingdom	3895	90.35
Germany	93	2.16
France	87	2.02
Spain	30	0.7
Belgium	25	0.58

90% of sales are from UK followed by Germany and France at 2% each.

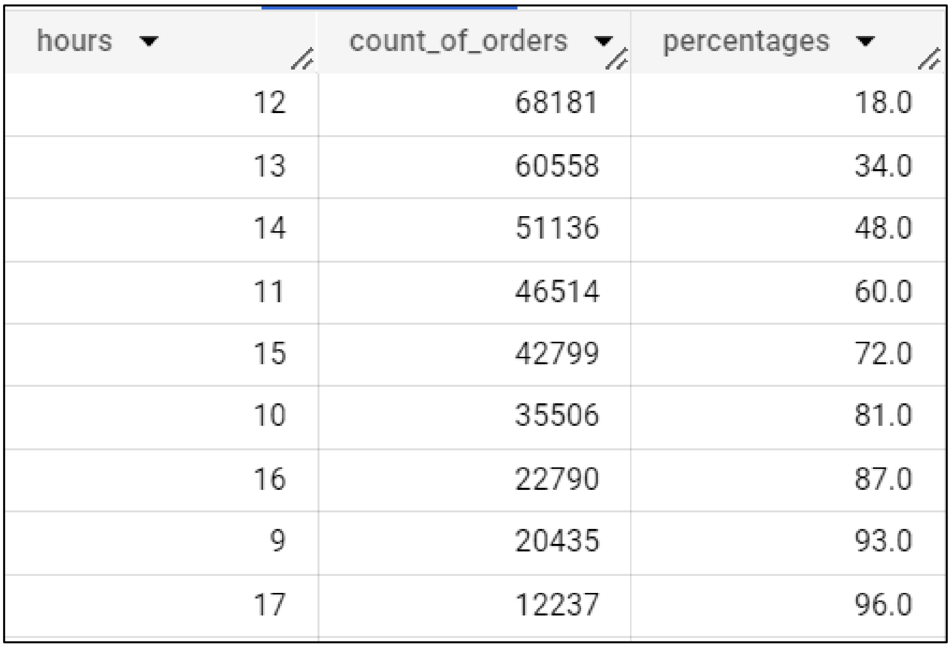
EXPLORATORY DATA ANALYSIS

Objective : Understand the peak hours of sales

QUERY 6

```
WITH cte as
  (SELECT
    *
  FROM
    (SELECT
      EXTRACT(HOUR FROM InvoiceDate) as hours,
      COUNT(*) as count_of_orders
    FROM `retail.sales`
    GROUP BY EXTRACT(HOUR FROM InvoiceDate)) a
  ORDER BY count_of_orders DESC)
SELECT
  *,
  ROUND((SUM(count_of_orders) OVER(ORDER BY
count_of_orders DESC )*100)/(SUM(count_of_orders)
OVER()),0) as percentages
FROM cte
ORDER BY count_of_orders DESC, percentages ASC
```

OUTPUT 6



hours	count_of_orders	percentages
12	68181	18.0
13	60558	34.0
14	51136	48.0
11	46514	60.0
15	42799	72.0
10	35506	81.0
16	22790	87.0
9	20435	93.0
17	12237	96.0

80% of sales are happening between 10 am to 3pm

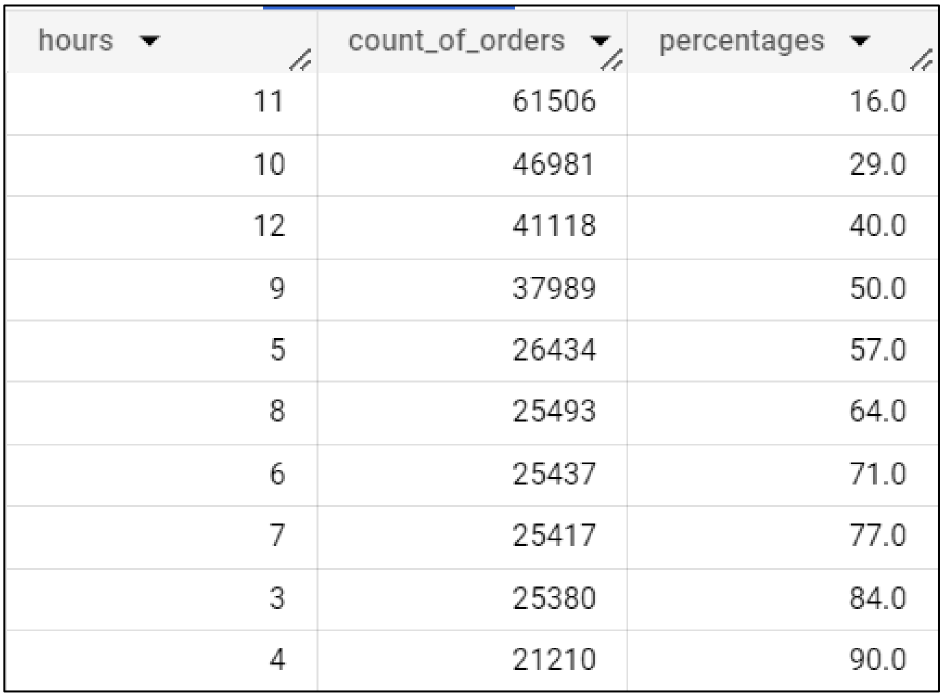
EXPLORATORY DATA ANALYSIS

Objective : Understand the peak months of sales

QUERY 7

```
WITH cte as
  (SELECT
    *
  FROM
    (SELECT
      EXTRACT(MONTH FROM InvoiceDate) as hours,
      COUNT(*) as count_of_orders
    FROM `retail.sales`
    GROUP BY EXTRACT(MONTH FROM InvoiceDate)) a
  ORDER BY count_of_orders DESC)
SELECT
  *,
  ROUND((SUM(count_of_orders) OVER(ORDER BY
count_of_orders DESC )*100)/(SUM(count_of_orders)
OVER()),0) as percentages
FROM cte
ORDER BY count_of_orders DESC, percentages ASC
```

OUTPUT 7



hours	count_of_orders	percentages
11	61506	16.0
10	46981	29.0
12	41118	40.0
9	37989	50.0
5	26434	57.0
8	25493	64.0
6	25437	71.0
7	25417	77.0
3	25380	84.0
4	21210	90.0

50% of sales are happening between September to December

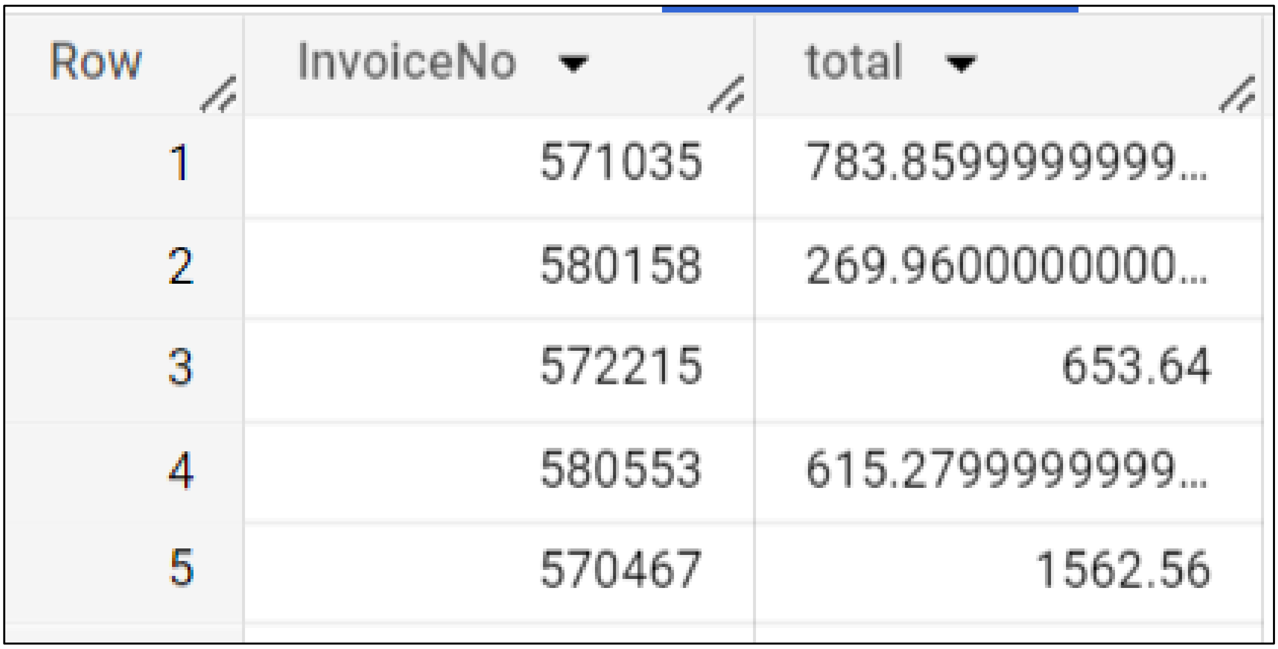
RFM ANALYSIS FOR CUSTOMER SEGMENTATION

Objective : Quantity bought and unit price is available in each row of the dataset. So the total cost corresponding to each invoice number is calculated using **inline calculations** in SQL i.e., Quantity * UnitPrice

QUERY 8

```sql
SELECT
 InvoiceNo,
 SUM(Quantity*UnitPrice) AS total
FROM
 `retail.sales`
GROUP BY
 InvoiceNo
```

OUTPUT 8



Row	InvoiceNo	total
1	571035	783.8599999999...
2	580158	269.9600000000...
3	572215	653.64
4	580553	615.2799999999...
5	570467	1562.56

The output is saved as a new table named bills in the database for future use

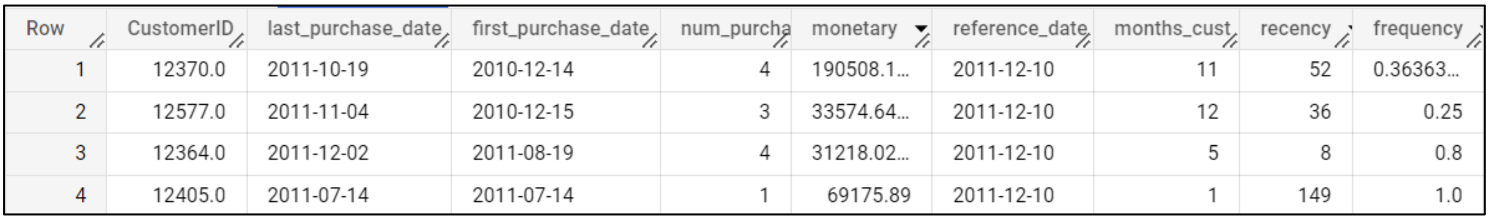
RFM ANALYSIS FOR CUSTOMER SEGMENTATION

Objective : Compute recency, frequency and monetary values for each customer
Recency : Reference date – last purchase date of each customer
(Reference date : Max (last purchase date in days of all the customers) + 1)
Frequency : No of purchases / (difference between first purchase date and last purchase date in months for each customer)
Monetary : Sum of the total purchase amount

QUERY 9

```
WITH cte AS
  (SELECT
    s.CustomerID,
    DATE(MAX(s.InvoiceDate)) AS last_purchase_date,
    DATE(MIN(s.InvoiceDate)) AS first_purchase_date,
    COUNT(DISTINCT s.InvoiceNo) AS num_purchases,
    SUM(b.total) AS monetary
  FROM
  `retail.sales`  s
  LEFT JOIN
  `retail.bills` b
  ON
    s.InvoiceNo=b.InvoiceNo
  GROUP BY
    CustomerID)
SELECT
  *,
  DATE_DIFF(reference_date, last_purchase_date, DAY) AS
recency,
  num_purchases/ (months_cust) AS frequency,
FROM
  (SELECT
    *,
    MAX(last_purchase_date) OVER () + 1 AS reference_date,
    DATE_DIFF(cte.last_purchase_date,
cte.first_purchase_date, month)+1 AS months_cust
  FROM cte)
```

OUTPUT 9



Row	CustomerID	last_purchase_date	first_purchase_date	num_purcha	monetary	reference_date	months_cust	recency	frequency
1	12370.0	2011-10-19	2010-12-14	4	190508.1...	2011-12-10	11	52	0.36363...
2	12577.0	2011-11-04	2010-12-15	3	33574.64...	2011-12-10	12	36	0.25
3	12364.0	2011-12-02	2011-08-19	4	31218.02...	2011-12-10	5	8	0.8
4	12405.0	2011-07-14	2011-07-14	1	69175.89	2011-12-10	1	149	1.0

The output is saved as a new table named RFM in the database for future use

RFM ANALYSIS FOR CUSTOMER SEGMENTATION

Objective : Group the customers into quintiles in terms of their RFM values and allot scores based on the table below

Percentile	Recency_Score r_score	Frequency_Score f_score	Monetary_Score m_score
0 – 20	5	1	1
20 - 40	4	2	2
40 - 60	3	3	3
60 - 80	2	4	4
80 - 100	1	5	5

Also ‘fm_score’ is calculated for the segmentation which is the average of f_score and m_score

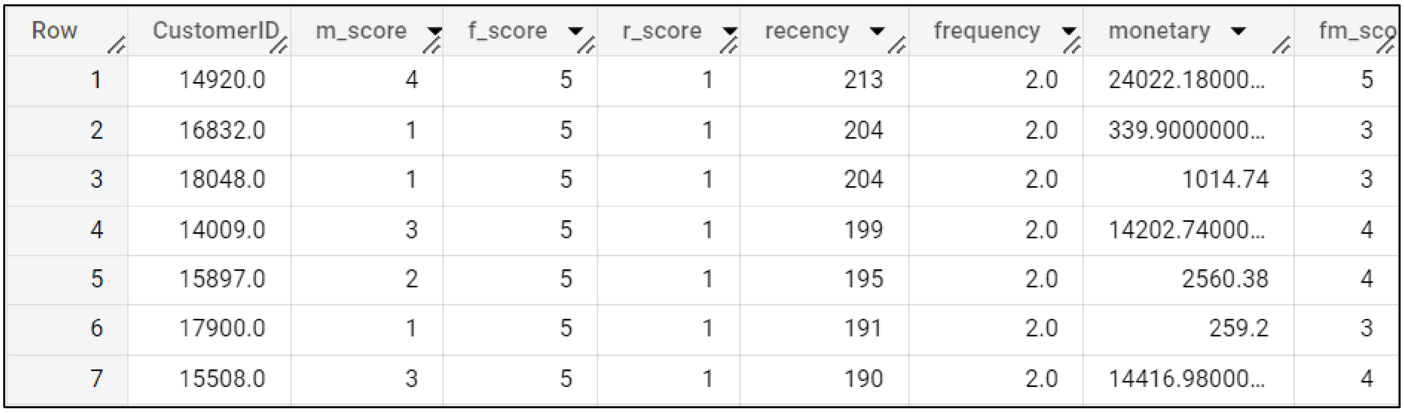
QUERY 10

```
SELECT
CustomerID,m_score,f_score,r_score,recency,frequency,monetary,
CAST(ROUND((f_score + m_score) / 2, 0) AS INT64) AS fm_score
FROM
 (SELECT *,
      --Monetary
  CASE WHEN monetary <= b.percentiles[(OFFSET(1))] THEN 1
  WHEN monetary <= b.percentiles[(OFFSET(2))] AND monetary >
b.percentiles[(OFFSET(1))] THEN 2
  WHEN monetary <= b.percentiles[(OFFSET(3))] AND monetary >
b.percentiles[(OFFSET(2))] THEN 3
  WHEN monetary <= b.percentiles[(OFFSET(4))] AND monetary >
b.percentiles[(OFFSET(3))] THEN 4
  WHEN monetary <= b.percentiles[(OFFSET(5))] AND monetary >
b.percentiles[(OFFSET(4))] THEN 5
  END AS m_score,
```

RFM ANALYSIS FOR CUSTOMER SEGMENTATION

```
    --Frequency
  CASE WHEN frequency <= c.percentiles[(OFFSET(1))] THEN 1
  WHEN frequency <= c.percentiles[(OFFSET(2))] AND
frequency > c.percentiles[(OFFSET(1))] THEN 2
  WHEN frequency <= c.percentiles[(OFFSET(3))] AND
frequency > c.percentiles[(OFFSET(2))] THEN 3
  WHEN frequency <= c.percentiles[(OFFSET(4))] AND
frequency > c.percentiles[(OFFSET(3))] THEN 4
  WHEN frequency <= c.percentiles[(OFFSET(5))] AND
frequency > c.percentiles[(OFFSET(4))] THEN 5
  END AS f_score,
    --Recency
  CASE WHEN recency <= d.percentiles[(OFFSET(1))] THEN 5
  WHEN recency <= d.percentiles[(OFFSET(2))] AND recency >
d.percentiles[(OFFSET(1))] THEN 4
  WHEN recency <= d.percentiles[(OFFSET(3))] AND recency >
d.percentiles[(OFFSET(2))] THEN 3
  WHEN recency <= d.percentiles[(OFFSET(4))] AND recency >
d.percentiles[(OFFSET(3))] THEN 2
  WHEN recency <= d.percentiles[(OFFSET(5))] AND recency >
d.percentiles[(OFFSET(4))] THEN 1
   END AS r_score,
   FROM
   `retail.RFM` a,
(SELECT APPROX_QUANTILES(monetary, 5) percentiles
FROM`retail.RFM`) b,
(SELECT APPROX_QUANTILES(frequency, 5) percentiles
FROM`retail.RFM`) c,
(SELECT APPROX_QUANTILES(recency, 5) percentiles
FROM`retail.RFM`) d)
```

OUTPUT 10

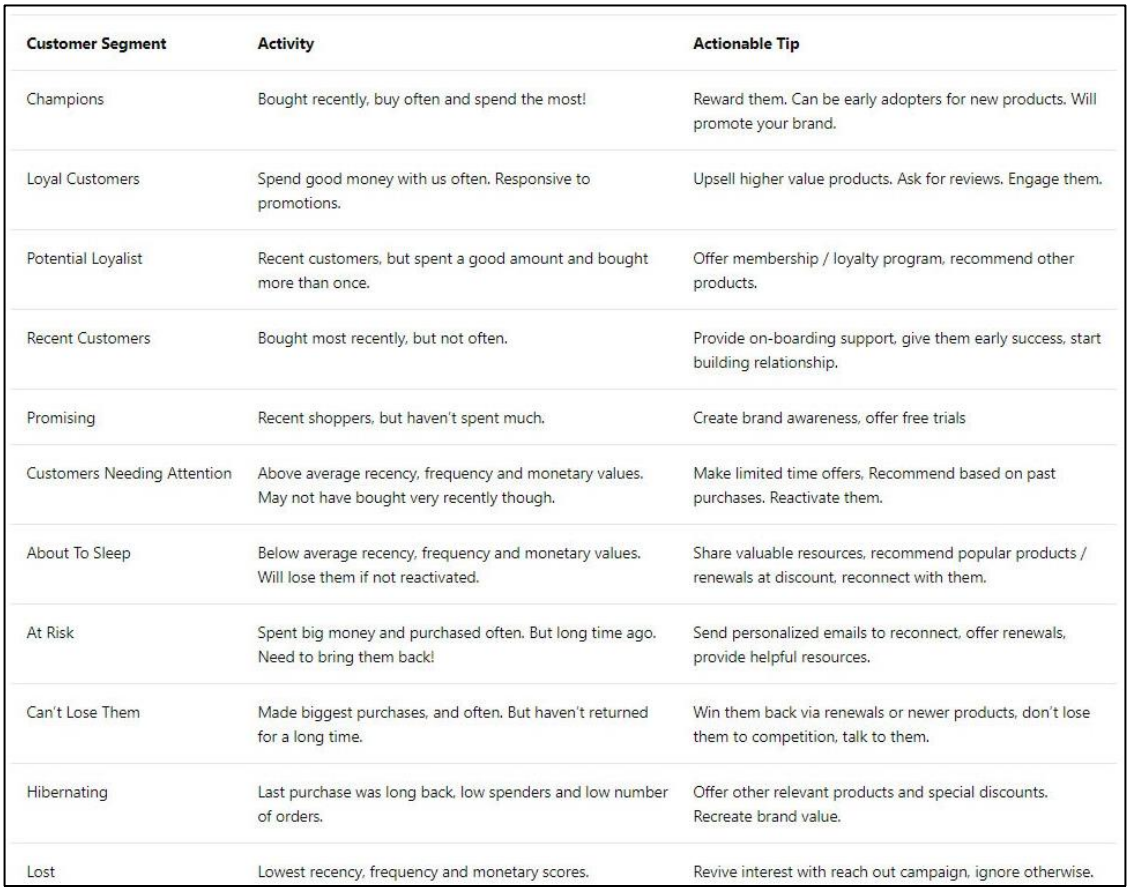


Row	CustomerID	m_score	f_score	r_score	recency	frequency	monetary	fm_sco
1	14920.0	4	5	1	213	2.0	24022.18000...	5
2	16832.0	1	5	1	204	2.0	339.9000000...	3
3	18048.0	1	5	1	204	2.0	1014.74	3
4	14009.0	3	5	1	199	2.0	14202.74000...	4
5	15897.0	2	5	1	195	2.0	2560.38	4
6	17900.0	1	5	1	191	2.0	259.2	3
7	15508.0	3	5	1	190	2.0	14416.98000...	4

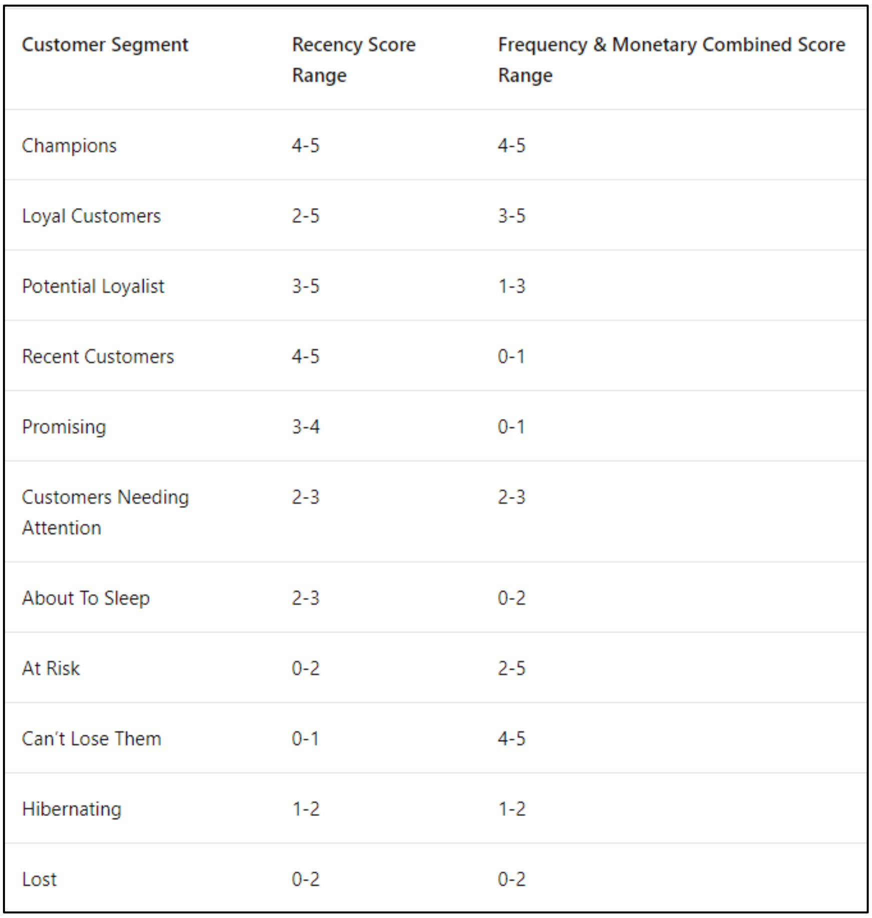
The output is saved as a new table named scores in the database for future use

RFM ANALYSIS FOR CUSTOMER SEGMENTATION

Objective : Group the customers into 11 personas (reference from UK Data & Marketing Association (DMA)) based on the RFM scores



Customer Segment	Activity	Actionable Tip
Champions	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new products. Will promote your brand.
Loyal Customers	Spend good money with us often. Responsive to promotions.	Upsell higher value products. Ask for reviews. Engage them.
Potential Loyalist	Recent customers, but spent a good amount and bought more than once.	Offer membership / loyalty program, recommend other products.
Recent Customers	Bought most recently, but not often.	Provide on-boarding support, give them early success, start building relationship.
Promising	Recent shoppers, but haven't spent much.	Create brand awareness, offer free trials
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.	Make limited time offers, Recommend based on past purchases. Reactivate them.
About To Sleep	Below average recency, frequency and monetary values. Will lose them if not reactivated.	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
At Risk	Spent big money and purchased often. But long time ago. Need to bring them back!	Send personalized emails to reconnect, offer renewals, provide helpful resources.
Can't Lose Them	Made biggest purchases, and often. But haven't returned for a long time.	Win them back via renewals or newer products, don't lose them to competition, talk to them.
Hibernating	Last purchase was long back, low spenders and low number of orders.	Offer other relevant products and special discounts. Recreate brand value.
Lost	Lowest recency, frequency and monetary scores.	Revive interest with reach out campaign, ignore otherwise.



Customer Segment	Recency Score Range	Frequency & Monetary Combined Score Range
Champions	4-5	4-5
Loyal Customers	2-5	3-5
Potential Loyalist	3-5	1-3
Recent Customers	4-5	0-1
Promising	3-4	0-1
Customers Needing Attention	2-3	2-3
About To Sleep	2-3	0-2
At Risk	0-2	2-5
Can't Lose Them	0-1	4-5
Hibernating	1-2	1-2
Lost	0-2	0-2

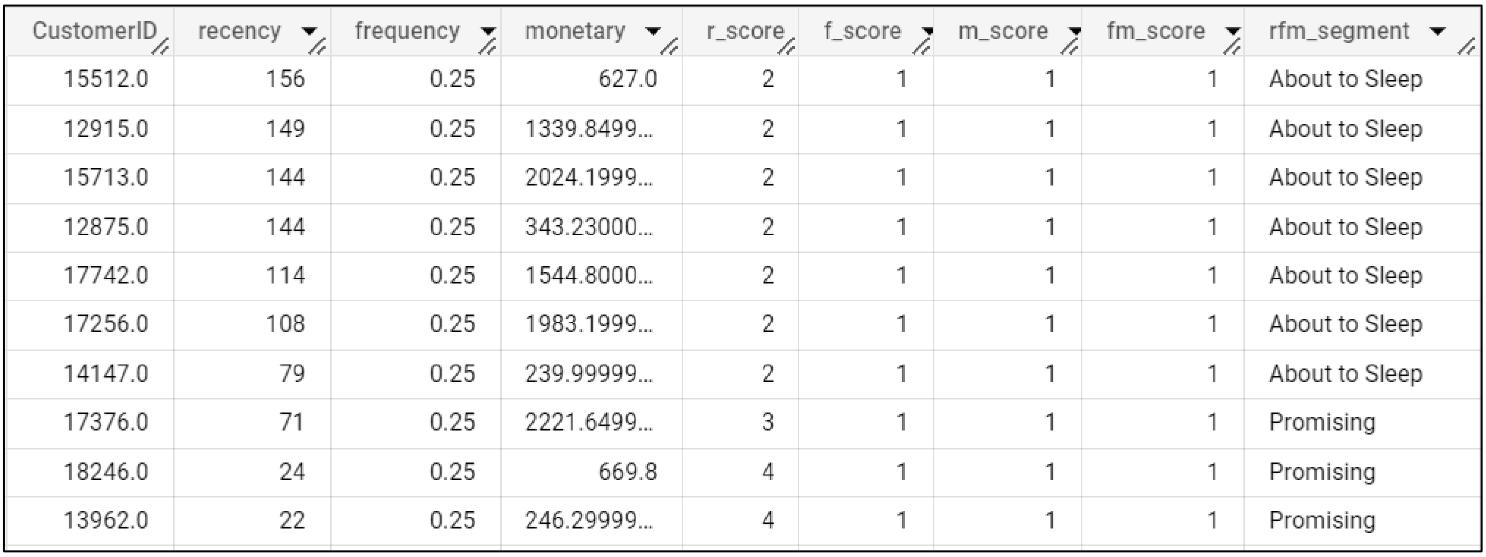
RFM ANALYSIS FOR CUSTOMER SEGMENTATION

QUERY 11

```
SELECT
    CustomerID,
    recency,frequency,monetary,
    r_score, f_score, m_score,
    fm_score,
    CASE WHEN (r_score = 5 AND fm_score = 5)
      OR (r_score = 5 AND fm_score = 4)
      OR (r_score = 4 AND fm_score = 5)
                                    THEN 'Champions'
    WHEN (r_score = 5 AND fm_score =3)
      OR (r_score = 4 AND fm_score = 4)
      OR (r_score = 3 AND fm_score = 5)
      OR (r_score = 3 AND fm_score = 4)
                                    THEN 'Loyal Customers'
    WHEN (r_score = 5 AND fm_score = 2)
      OR (r_score = 4 AND fm_score = 2)
      OR (r_score = 3 AND fm_score = 3)
      OR (r_score = 4 AND fm_score = 3)
                                    THEN 'Potential Loyalists'
    WHEN r_score = 5 AND fm_score = 1 THEN 'Recent Customers'
    WHEN (r_score = 4 AND fm_score = 1)
      OR (r_score = 3 AND fm_score = 1)
                                    THEN 'Promising'
    WHEN (r_score = 3 AND fm_score = 2)
      OR (r_score = 2 AND fm_score = 3)
      OR (r_score = 2 AND fm_score = 2)
                                    THEN 'Customers Needing Attention'
    WHEN r_score = 2 AND fm_score = 1 THEN 'About to Sleep'
    WHEN (r_score = 2 AND fm_score = 5)
      OR (r_score = 2 AND fm_score = 4)
      OR (r_score = 1 AND fm_score = 3)
                                    THEN 'At Risk'
    WHEN (r_score = 1 AND fm_score = 5)
      OR (r_score = 1 AND fm_score = 4)
                                    THEN 'Cant Lose Them'
    WHEN r_score = 1 AND fm_score = 2 THEN 'Hibernating'
    WHEN r_score = 1 AND fm_score = 1 THEN 'Lost'
    END AS rfm_segment
  FROM `retail.scores`
```

RFM ANALYSIS FOR CUSTOMER SEGMENTATION

OUTPUT 11



CustomerID	recency	frequency	monetary	r_score	f_score	m_score	fm_score	rfm_segment
15512.0	156	0.25	627.0	2	1	1	1	About to Sleep
12915.0	149	0.25	1339.8499...	2	1	1	1	About to Sleep
15713.0	144	0.25	2024.1999...	2	1	1	1	About to Sleep
12875.0	144	0.25	343.23000...	2	1	1	1	About to Sleep
17742.0	114	0.25	1544.8000...	2	1	1	1	About to Sleep
17256.0	108	0.25	1983.1999...	2	1	1	1	About to Sleep
14147.0	79	0.25	239.99999...	2	1	1	1	About to Sleep
17376.0	71	0.25	2221.6499...	3	1	1	1	Promising
18246.0	24	0.25	669.8	4	1	1	1	Promising
13962.0	22	0.25	246.29999...	4	1	1	1	Promising

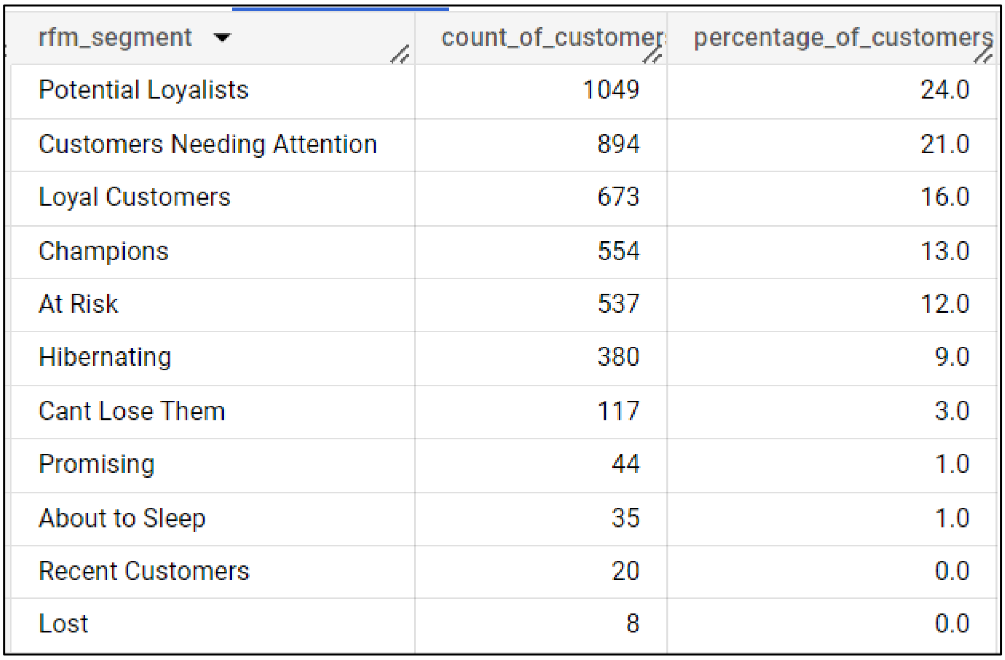
RFM ANALYSIS FOR CUSTOMER SEGMENTATION

Objective : Percentage of customers corresponding to each RFM segment

QUERY 12

```
 WITH cte as
  (SELECT
    rfm_segment,
    COUNT(DISTINCT CustomerID) as count_of_customers
  FROM `retail.rfm_divided`
  GROUP BY rfm_segment),
  cte1 as
  (SELECT
    COUNT(DISTINCT CustomerID) as total_count
  FROM `retail.rfm_divided`)
SELECT
  c.rfm_segment,
  c.count_of_customers,
  ROUND(c.count_of_customers*100/c1.total_count,0) as
percentage_of_customers
FROM cte c, cte1 c1
ORDER BY c.count_of_customers DESC
```

OUTPUT 12



rfm_segment	count_of_customer	percentage_of_customers
Potential Loyalists	1049	24.0
Customers Needing Attention	894	21.0
Loyal Customers	673	16.0
Champions	554	13.0
At Risk	537	12.0
Hibernating	380	9.0
Cant Lose Them	117	3.0
Promising	44	1.0
About to Sleep	35	1.0
Recent Customers	20	0.0
Lost	8	0.0

INSIGHTS

1. As per the data between 2010-2011, the retail store sold more than 3600 distinct products to 4300+ customers across 37 different countries.
2. The unit price of top 5 revenue contributing products are less than or equal to the average unit price considering all the products in the retail store. This is obvious because lesser the price, more will be the sales of the product.
3. 1700 products accounts for 80% of the sales
4. 90% of sales are from UK followed by Germany and France at 2% each.
5. 80% of sales are happening between 10 am to 3pm.
6. 50% of sales are happening between September to December nearing the Christmas & New Year.
7. % of customers 'lost' are almost nil which is a good sign for the business showing excellent customer retention.
8. More than 50% of the customers are 'Potential loyalists' , 'Customers needing attention' and 'Loyal customers'.
9. 13% of the customers are the 'Champions' whos spends the most and has high recency and frequency score.
10. At the same time there are 12% customers who are at risk of moving into danger zone with below average spendings, frequency and recency score who needs attention.

RECOMMENDATIONS

1. Although there are sales in 37 different countries, 90% accounts to UK and Germany and France are at 2% each. This is a huge difference and so there is a need to do an in depth analysis country wise to promote the sales in different countries apart from UK. Since the sales are high in UK, the marketing campaigns should be focused on other countries including Germany & France.
2. 1700 out of 4300+ products accounts for 80% of the sales. Some of the low selling products can be stopped which will reduce the inventory requirements while maintaining the revenue at similar levels.
3. Large number of stocks to be maintained during the year end between September to December which are the months of peak sales in a year.
4. Advantages can be given to 'Champion' customers on new product releases or flash sales with early intimations and reminders, initial booking advantages for new product releases.
5. Credit card facilities can be given to 'champions' and 'loyal customers' at reduced interest rates as a reward for high RFM scores and also to increase customer retention
6. Those customers at risk of losing (who used to be frequent buyers but the recency score is very low) can be reconnected with personalized emails, discount vouchers on products they used to buy earlier.
7. Those coming under the 'lost' segment with low RFM scores can be ignored because the percentage is very low to invest resources and since the monetory value is also low, the value addition with respect to revenue will be very low