

## Performing an RFM analysis with SQL and Excel

Data Source:

- <https://www.kaggle.com/datasets/ylchang/coffee-shop-sample-data-1113>

Technologies:

- Postgres SQL 16
- Excel

While performing an exploratory data analysis on our sales data from April, we have discovered a large decline in purchases from customers registered in our loyalty program.

We are interested in identifying customers that are able to be enticed back with a marketing program; we will be searching for customers that have not returned recently.

To do so, we will run an RFM Analysis to establish customer segmentation with recency being our largest focus.

## Part 1, Establishing RFM Data in PostgreSQL databases

### Creating Order\_ID column

/\*

Before we can begin RFM analysis, we need to create a unique order column in Receipts. This is because transaction\_id does not uniquely identify the order; transaction\_id repeats per location and per day.

To create a unique order\_id column, we will merge transaction\_id, transaction\_date, and sales\_outlet\_id to create a unique column per order which will provide accurate count to the number, recency, and total value of orders.

\*/

--Create our "Order\_ID" Column:

```
ALTER TABLE RECEIPTS ADD COLUMN ORDER_ID TEXT;
```

-- Verify Creation:

```
SELECT *
```

```
FROM RECEIPTS
```

```
LIMIT 5
```

-- Add concat values to new column 'order\_id' by update, set using concatenation function

```
UPDATE RECEIPTS
```

```
SET ORDER_ID = CONCAT(TRANSACTION_DATE,TRANSACTION_ID,SALES_OUTLET_ID)
```

-- verify results, count distinct should result in different number than number of composite parts

```
SELECT COUNT(DISTINCT TRANSACTION_ID) AS TRANSIDCOUNT,
```

```
       COUNT(DISTINCT TRANSACTION_DATE) AS DATECOUNT,
```

```
       COUNT(DISTINCT ORDER_ID) AS ORDERIDCOUNT
```

```
FROM RECEIPTS
```

### Creating Query to export to CSV

/\* Now confirmed, staging into new table and creation of a CSV file for RFM Analysis should occur.

The necessary components we want from our customer group will be:

Recency: Max transaction\_date

Frequency: count(distinct order\_id)

Monetary: sum(line\_item\_amount)

We will include two additional columns:

Customer\_ID

Average order value = sum(line\_item\_amount)/count(distinct order\_id)

Average order value is included

It should be noted here that we will exclude customer ID = 0 as this is an unregistered customer.

We can compare, at a later date, the value of orders of registered vs unregistered to identify effectiveness of loyalty program spend.

\*/

COPY (

```
SELECT MIN(CUSTOMER_ID) CUSTOMER_ID,
       COUNT(DISTINCT ORDER_ID) ORDER_COUNT,
       MAX(TRANSACTION_DATE) LATEST_PURCHASE,
       SUM(LINE_ITEM_AMOUNT) TOTAL_REV,
       (SUM(LINE_ITEM_AMOUNT)/COUNT(ORDER_ID)) AVERAGE_ORDER_VALUE
FROM RECEIPTS
WHERE CUSTOMER_ID <> '0'
GROUP BY CUSTOMER_ID
ORDER BY TOTAL_REV DESC
```

)

TO 'D:\DATA\COFFEE SHOP SAMPLE DATA\RFM\_DATA.CSV' DELIMITER ',' CSV HEADER;

## Part 2: Excel Data Analysis

We will have a CSV document with the below columns that will allow us to begin ranking.

customer_id	order_count	latest_purchase	total_rev	average_order_value
-------------	-------------	-----------------	-----------	---------------------

For this analysis, we will be using quartiles to determine what categories the customer fits within.

We will also use 5/1/2019 (the first day of the next month) as the end date to determine # of days for recency.

Finally, we will take advantage of the “Quartile.inc” function in excel to determine bounds on the quartiles with the “IFS” condition measuring against the quartiles.

Using 5/1/2019. We are able to define an end date, and a number of days since last order column.

days since last purchase	R Rank	F Rank	M Rank	RFM Ranks:	At Risk Column
--------------------------	--------	--------	--------	------------	----------------

R-Rank, F-Rank, and M-Rank are defined by the quartile they fit in, and we are able to conduct our analysis.

We are able to take advantage of a Pivot table to filter down to the “At Risk Column”, 557 of 2247 total customers (just shy of 24.68%).

With the list of customers who are “At risk” compiled, we were able to take a look at a few other options on why we may be seeing a higher-than-normal at-risk customer count.

We have taken a look at 3 factors that may indicate cause of loss to prevent or more specifically target to stop future loss. Those factors are:

In addition to our RFM Analysis, we found some interesting Demographic information utilizing pivot tables and pivot charts.

We have uncovered that there is currently a generational shift in New customer registrations from GenX, older Millennials, and Baby Boomers to GenZ. This shift was not significant enough to offset total registrations for April.

