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:

Sign up Year

Default Store

Customer Age

The results are below:

Year:	At risk	Population	%Diff
2017	235	986	-1.63%
2018	253	988	1.51%
2019	68	272	0.12%

Store	At risk	Population	% diff
3	240	800	7.55%
5	228	945	-1.07%
8	88	501	-6.48%

Customer Age Bucket	At risk	Population	% diff
18-30	185	790	-1.90%
31-40	125	470	1.56%
41-50	99	348	2.31%
51-60	72	329	-1.70%
61-70	75	309	-0.27%

% diff:

(at_risk_segment/total_at_risk)- (population_segment/total_population)
We are looking for large differences between the at-risk group and population as an indication to an issue

Store may be worth looking into more deeply, as store 3 appears to have a larger percent of at-risk customers than as a % of population would be.

Part 3: Recommendation

It is current recommendation to search more deeply on cause of the reduction of sales from customers enrolled in our loyalty-program. This will be a worthwhile analysis to re-compute for the following months to attempt to identify a trend.

At current, Store may be a factor, so it is worth keeping an eye on this.

Finally, we are able to derive the customer set with email and name to create a marketing promotional effort. It is recommended to potentially include a more aggressive promotional structure for loyalty members both as a means to re-engage “at-risk” customers.