Segmentation of Gate City Bank's customer using RFM analysis (2016-2022)

Trang Nguyen

A Project Submitted in Partial Fulfillment of the Requirements of

the Master of Science in Business Analytics

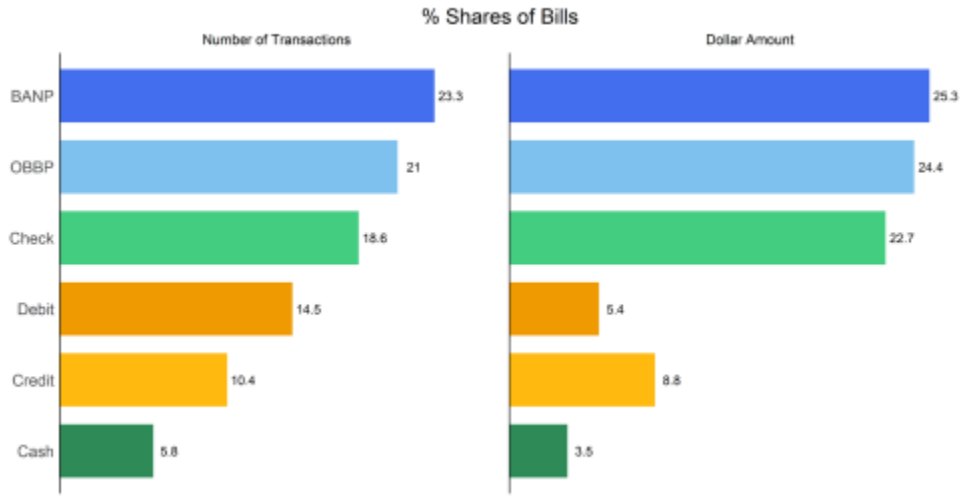
North Dakota, May 2023

I. Introduction

1. Business information:

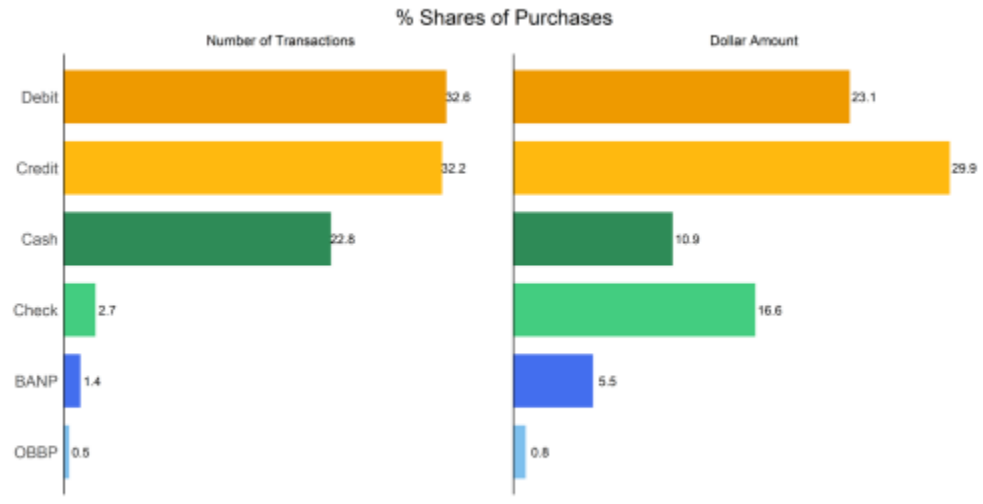
According to a research of 2020 Diary of Consumer Payment Choice, Debit cards accounted for almost 33% of transaction purchases and 14.5% of bill payment (as images below) [1]. Therefore, debit cards play an important role in economics. It can help the financial institution evaluate the value of customers regarding customers' account balance, the fee collected from the transaction customers paid for.

Figure 5: Payment instrument use for bills, shares by number and value



Note: Shares do not sum to 100 because less common payment instruments are omitted.
Source: 2020 Diary of Consumer Payment Choice

Figure 6: Payment instrument use for purchases, shares by number and value



Note: Shares do not sum to 100 because less common payment instruments are omitted.
Source: 2020 Diary of Consumer Payment Choice

Gate City Bank is the region's leading financial institution whose headquarters is in Fargo, North Dakota. They have 43 locations in 22 communities across North Dakota and central Minnesota. Beside offering the credit and lending services, which made them well known for, Gate City Bank also issues debit cards for customers, which can be referred to non-contractual customers. The risk

debit cards bring to the institution is less than one from credit cards, but much profit. However, that does not mean that the information from customer's spending is not valuable. The financial institution would like to explore the debit cards spending of individuals' data combining with customers' demographic information to create and foster innovation to support strategic goals.

2. Business task:

The purpose was to segment customers based on their spending behaviors and income which could give the financial institution more information about their clients to plan data-driven strategies for those who are using the debit card product. The key insights drawn from the analysis will help improve the bank promotion for maintaining their loyal customers, reactivating dormant customers based on their card usage.

II. Data collection and preparation:

1. Data sources:

There is a big folder for three objects: Customer, Account and Transaction from 2016 to 2023 in the dataset. There is a problem with a dataset that is big. Descriptions for each variable can be found in the appendix.

- Customer: 2.5 GB (27 CSV files)
- Account: 5.5 GB (56 CSV files)
- Transaction: 50 GB (496 CSV files)

Generally, one household key can have multiple customers, one customer can have multiple accounts, and one account can have multiple transactions with their cards. However, the bank adds another dimension which is eom_prod_dt. This dimension is the date at the end of each month. We always have to consider the time dimension when processing data to have accurate information. If not, there will be a mess up between household_key and account_key. The relationship between each tables are described below (Figure 1), with an assumption that the eom_prod_dt is fixed at Dec 31st,2022:

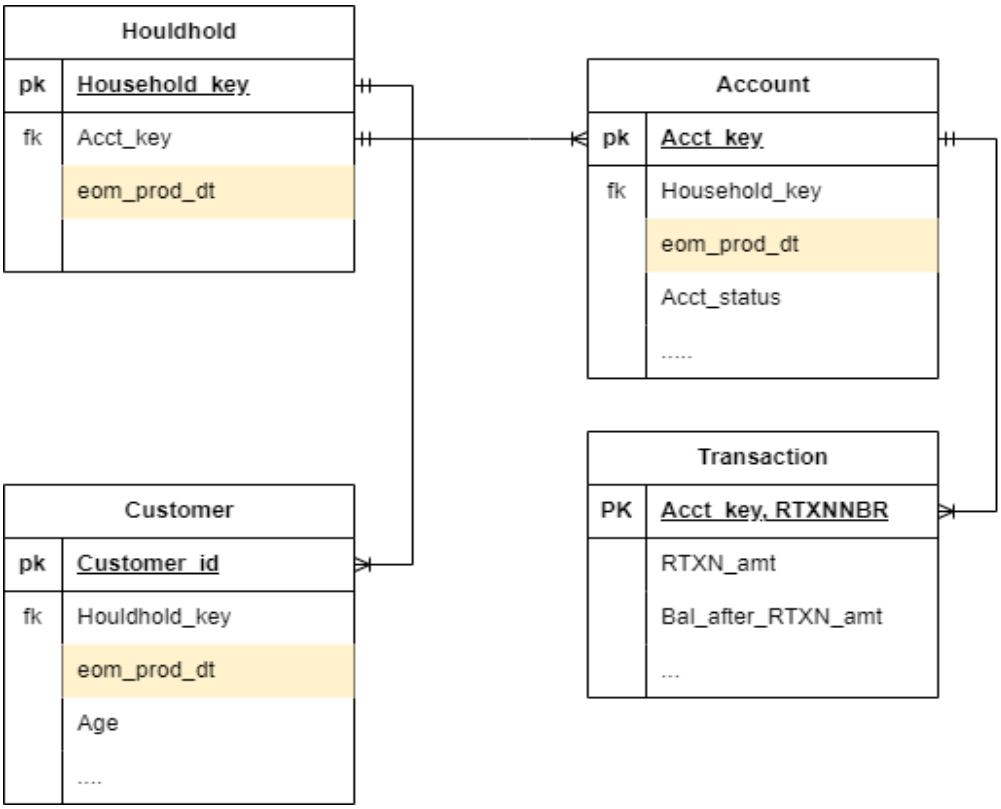


Figure 1: Diagram of database structure

To get more information, we chose to evaluate at the account level. For demographic data, we only have Age; there is no clear connection between customers and accounts other than the household key. Also, the RFM model requires data on bank customers' spending solely to segment and examine their trends. As a result, we mostly concentrate on the transaction tables.

2. Read dataset CSVs into dataframes:

We use both PySpark and Pandas to merge and filter tables together for all objects (Customers, Account, Transactions) depending on their size.

Customer:

379,836 distinct rows which is equivalent to the number of all their customers during the time; including business and personal customers.

Account:

Before filtering, we had 698,054 accounts with both personal and business accounts, all account statuses (active, close, dormant,...), and all account categories (checking, savings, mortgage,...). There was a system migration in 2016, which resulted in a difference in the number of accounts for the entire time compared to the number of accounts at the end of 2022, 570,987 accounts.

Total rows: 34,068,363:

- Distinct account for all categories: 698,054 acc

```
count_dis = df.groupBy(F.col("account_key")).agg(F.countDistinct(F.col("account_key")))
count_dis.show()
```

```
[16] count_dis.count()

     698054
```

- Eom_prod_dt fixed at 12/31/2022: 570,987 = > assumption: the different is there will be more accounts open after that

```
[26] df_date = df.filter(df.eom_prod_dt == "2022-12-31")
     df_date.count()

     570987
```

Time dimension is a crucial component for this dataset, as we have already stated. Hence, in order to obtain the specific number at that time, I fixed the variable eom_prod_dt to 12/31/2022.

We used filters to obtain data for checking and personal accounts only, and as of 12/31/2022, we had 205,396 accounts in total.

account_key	ACTIVE_DATE	account_s	closed_dt	eom_prod_dt	balance	orig_bal_a	major_typ	major_de	minor_description	customer_	Household_key
929139a7352b006950d848f36c64436b	8/7/2000	ACT		12/31/2022	4499.84	200	CK	Checking	Benefit Interest Checking	Personal	ca24e08f392533394fcd828530c0c7ac
8ad863c33665d152566c75ac0c6baf7c	12/6/2000	ACT		12/31/2022	783.64	300	CK	Checking	Benefit Interest Checking	Personal	0315ef8cd2d8540e505acb8e241493ce
853c8bb83f1cc410ca9f47cf30e92ed0	10/1/2003	ACT		12/31/2022	889.13	252.27	CK	Checking	Benefit Interest Checking	Personal	01a91ece570294b158cb161a9ac5af8f
50e4234ab57bfbe25f5ac69e8c50cb8a	3/19/2004	ACT		12/31/2022	33920.43	50	CK	Checking	Benefit Interest Checking	Personal	d38d93d5eec2f3a02e0ae6e7b2e36f6f
564b163c3c0d5d6cfc77bafda824bd29	4/13/2004	ACT		12/31/2022	980.64	300	CK	Checking	Benefit Interest Checking	Personal	7b64a70e780fb89d6cd3b37b76b35327
c986f335be1ab539a64981f8cb740327	12/30/2004	ACT		12/31/2022	20569.05	124.63	CK	Checking	VIP FREE Interest Checking	Personal	0af74d5d4f419b2858c2905d7c0912ec
f4541eff496edc00f73c81e79f38b2dc	6/13/2005	ACT		12/31/2022	1336.83	107	CK	Checking	Benefit Interest Checking	Personal	49ddb93a666036eaa3bbe10fcf13f000
1a4542f2105a0d664d8f4c5de3db0bdd	3/12/2005	ACT		12/31/2022	30538.48	2010	CK	Checking	VIP FREE Interest Checking	Personal	d5a40eb9eafc7982b056eed916c966a0
11a868b3690787ec1e8de72de5027749	3/19/2005	ACT		12/31/2022	1927.06	92.95	CK	Checking	Benefit Interest Checking	Personal	da49c546b6b90335ccdbc22ac0d67b7f
abc11069699121f0d1061a9a5ac32624	10/18/2005	ACT		12/31/2022	456.86	200	CK	Checking	Benefit Interest Checking	Personal	243f9ee53bb5991497294b894407d7af
76c59c2ad6f958e254d6af712a37519e	4/14/2006	ACT		12/31/2022	22468.28	510	CK	Checking	Benefit Interest Checking	Personal	536b6ebe20b8d303c16452f29c991b58
8abbf263513f110b1dabb140ba23e6da	1/22/2007	ACT		12/31/2022	15.75	90	CK	Checking	Benefit Interest Checking	Personal	bae2e1631726169b5f9632c966f02b01
9c41bab53b028daa461c9c423bd66961	8/13/2007	ACT		12/31/2022	1702.91	450	CK	Checking	Totally FREE Checking	Personal	f4396778bb66b423b7ad8ed52be9c04f
6fa5e381214635f7c73c256e4fcb27bb	1/12/2008	ACT		12/31/2022	106.99	302.8	CK	Checking	Totally FREE Checking	Personal	9898643c7ce281994344410cf6776cda
4c2baafcaf6289ed3ea12c829bf59a60	1/12/2008	CLS	########	12/31/2022	0	400	CK	Checking	Benefit Interest Checking	Personal	c165038ce6422b5b2fd1c3784e1c7737

Transaction:

The size of the raw data is 225,178,036 rows x 18 columns. We only choose the two types of transactions that have the most influence on the account, given the project's goals and scope.

- External deposits, positive value, are often the owner of the account's income or salary. Assuming that they have just one source of income, we can estimate their income for this transaction.
- Point-of-sale withdrawals, negative value, show how cardholders use their accounts.

TYPE OF TRANS	AMT	COUNT	PER_amt	PER_COUNT
XDEP	$ 13,102,953,430.72	11,451,624	29.62%	5.65%
DEP	$ 7,443,002,507.93	7,642,569	16.83%	3.77%
XWTH	$ (6,547,597,885.11)	21,580,713	14.80%	10.66%
PWTH	$ (4,966,069,233.74)	119,458,933	11.23%	58.99%
WTH	$ (4,956,852,941.68)	7,056,172	11.21%	3.48%
CWTH	$ (3,496,931,962.68)	6,024,780	7.91%	2.98%
DEPD	$ 1,105,482,099.12	1,497,693	2.50%	0.74%
DWTH	$ (725,166,621.17)	5,989,698	1.64%	2.96%

So, we used the filters listed below to reduce its size to a level that was suitable for our analysis.

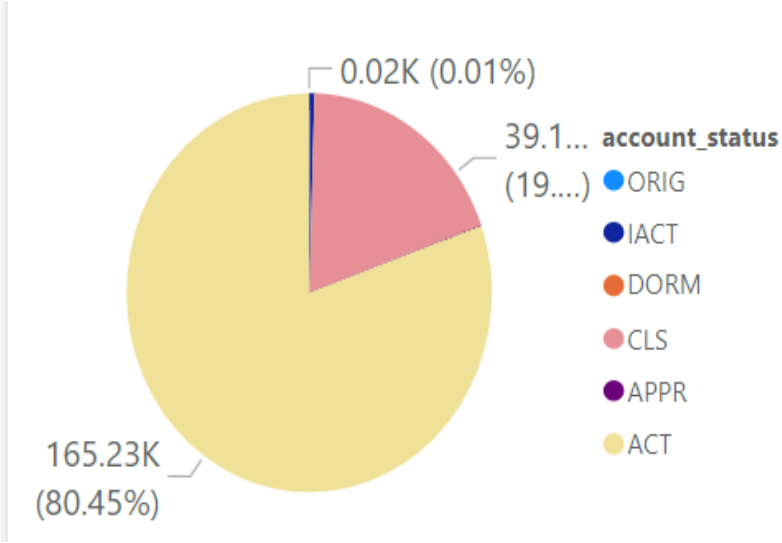
- Type of transaction: point of sale withdrawal, external deposit
- Categories of account: Checking and personal account
- Status of transaction: clear (RTXN = 1)

After all, the size reduces dramatically to around 130 million rows.

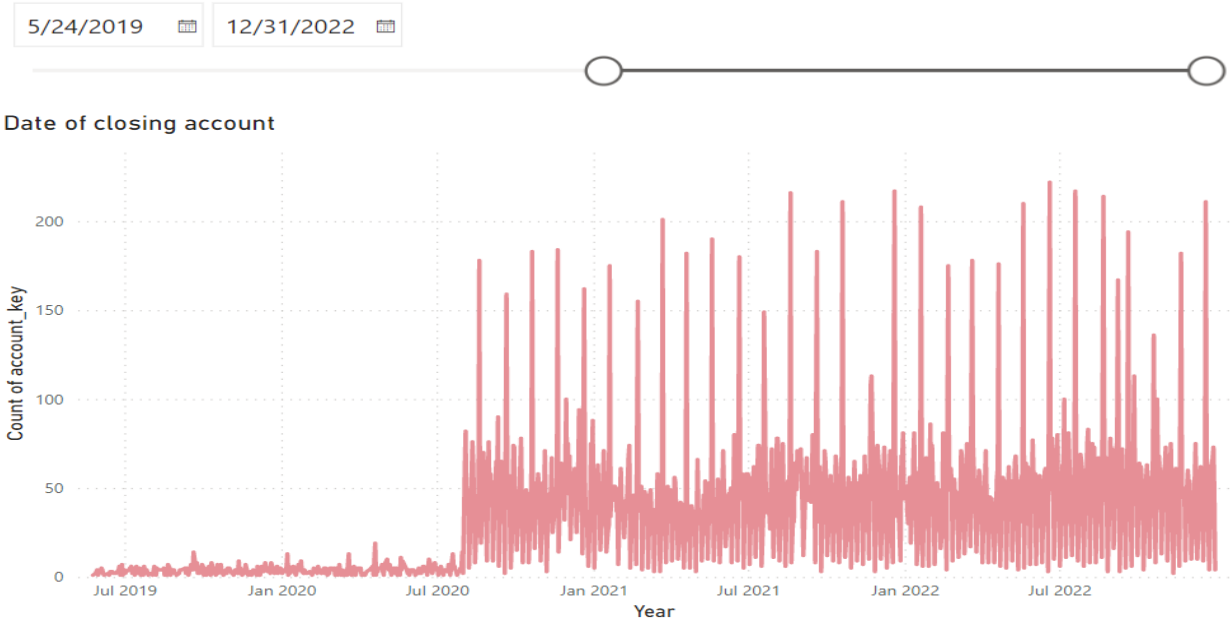
III. Visualization:

1. Account:

We have 80% of all personal checking accounts active, another approximately 20% are closed, and a very minor number are in any other status, with eom_prod_dt fixed at 12/31/2022.

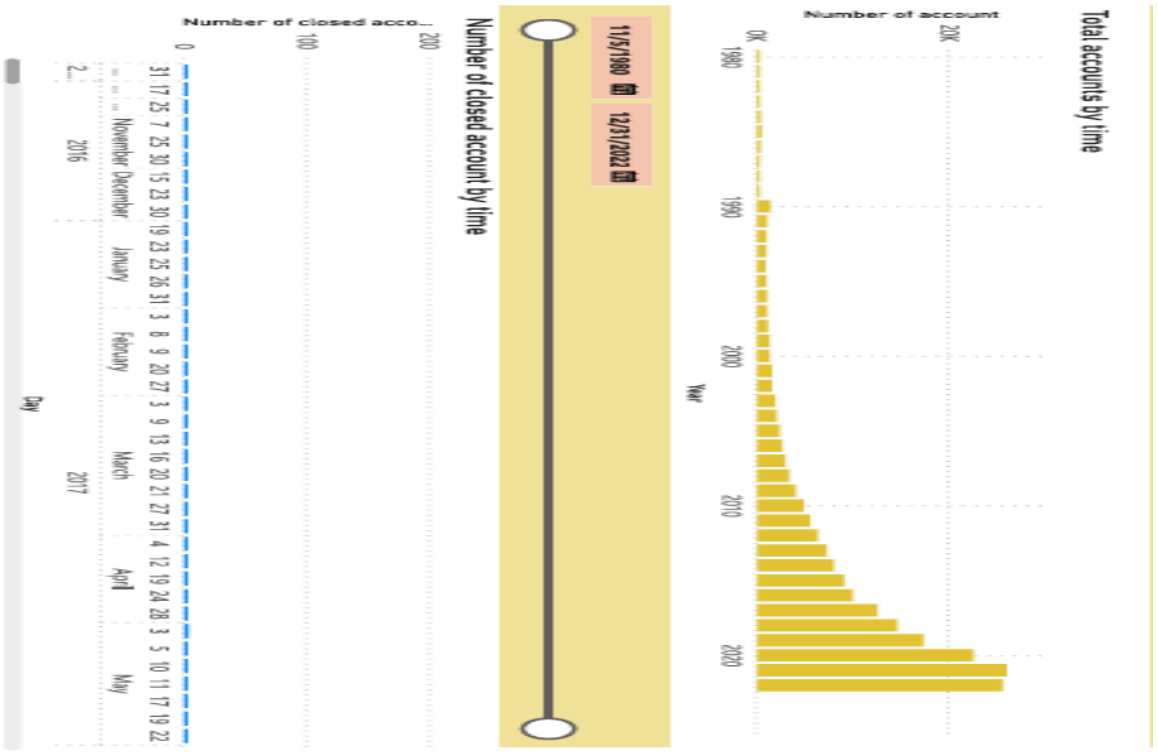


As of January 2021, there is a significant change in the number of accounts closing. The number is particularly highest towards the middle of the month. After that, the pattern, low at the beginning and high in the middle of the month, repeats itself and remains reliable.



orig_bal_amt (bins)	Count of account_key
0.00	17575
100.00	4936
200.00	2644
300.00	1651
400.00	1217
500.00	1644
600.00	758
700.00	588
800.00	475
900.00	348
1,000.00	990
Total	**38119**

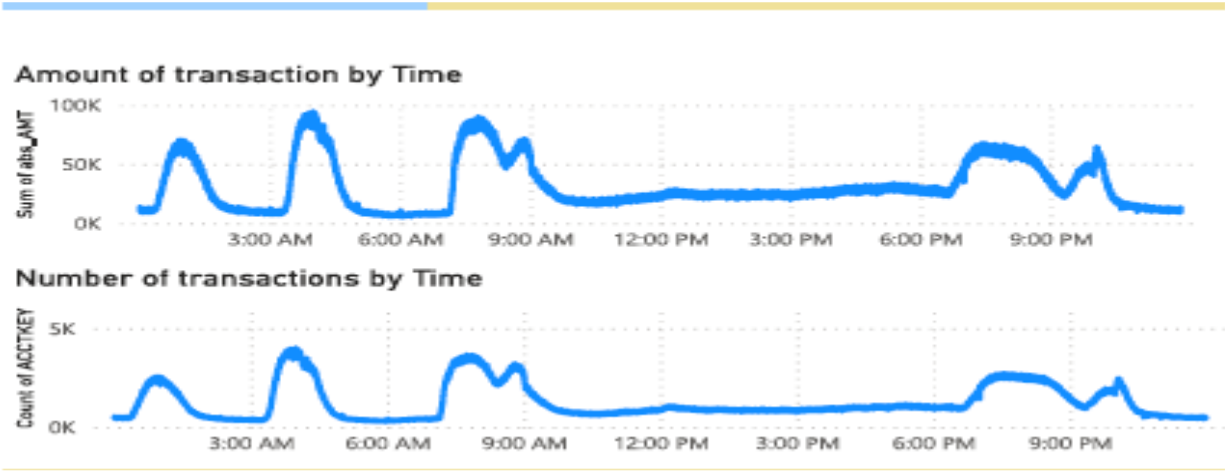
Over 50% of new accounts are opened with a deposit of less than $100, and just 16% of customers have deposits of more than $1,000. The $1000 deposit mark serves as the critical threshold since clients are less likely to close their accounts if they deposit more money into their accounts for the first time.

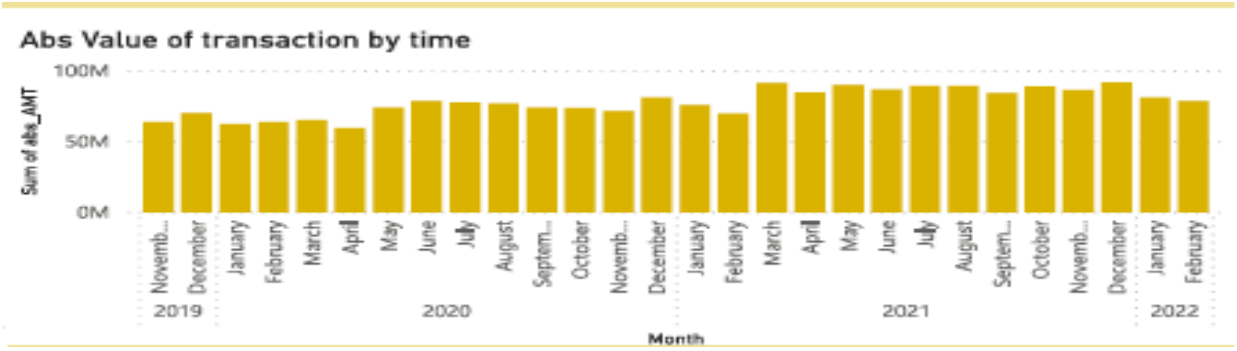


In general, the number of accounts has increased over time, from 1980 to the present, and will continue to rise after 2020. The bank will automatically close the account in IACT status in the middle of the month.

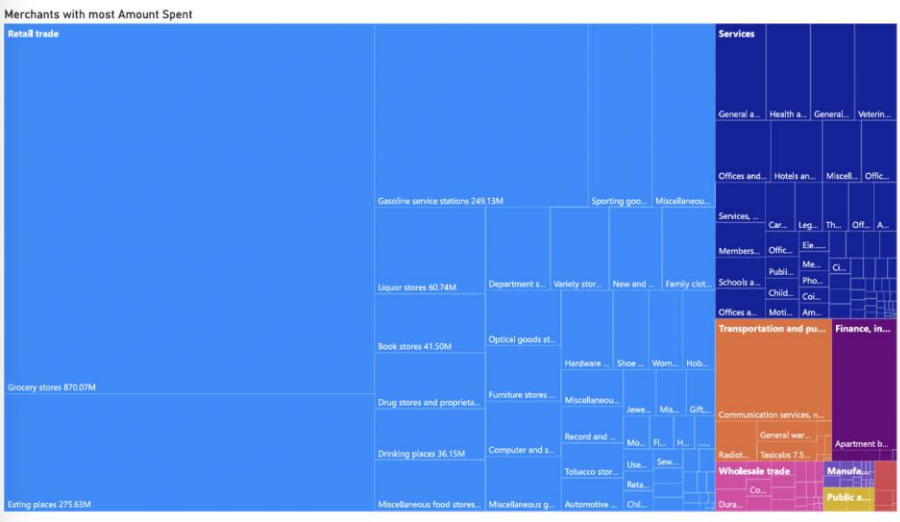
2. Transaction:

The graph above describes the times that customers used their cards the most. After midnight, the number of transactions and total amount of transactions from 12:00 AM - 2:00 AM is quite high.

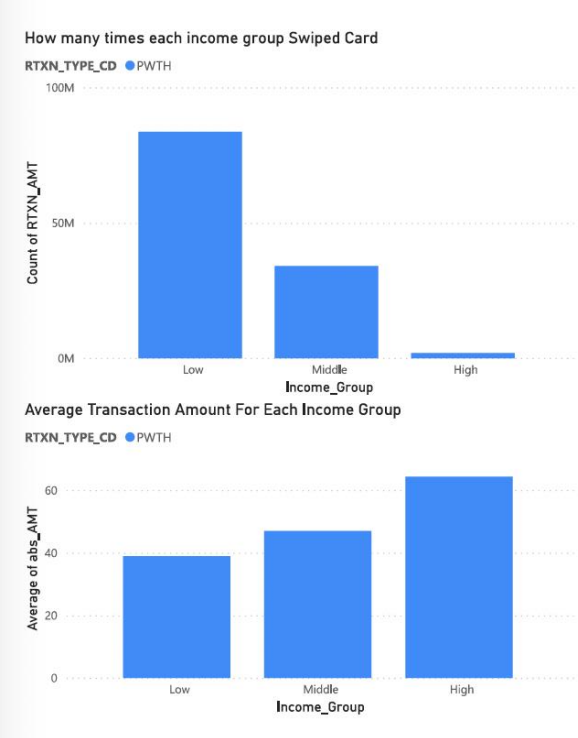




Each year, the columns of March and December are higher than the others because it is tax season and the Christmas season. That trend is being followed by both spending and income.



We tried to look at what kind of merchants customers most spend at. We retail trade receives the most amount of spend followed by services. Within retail trade Grocery Stores are at the top followed by Gas Stations and then Restaurants or Eating places. We can use this data to determine what sort of merchants should Gate City partner with in order to give their customers promotions and/or discounts.



We grouped customers using their income group based on these incomes:

Low income	Less than $52,200
Middle income	$52,200 - $156,600

Upper income	More than $156,600

We see that the high income group swipes the card the least while low income swipes their debit card the highest. It could be that people with high income usually have higher credit scores which could mean that they use credit cards because they get points and discounts through them. Low income people might have low credit scores which is why they have to use their debit cards.

We looked at the average transaction amount for each income group and it seems that the high income group has a higher avg. transaction amount followed by middle and then low income groups. This shows us that high income groups tend to buy more costly goods/services than low income customers.

IV. Model

1. RFM Score:

1.1. What is RFM:

RFM stands for Recency, Frequency, and Monetary Value, which is a marketing analytics framework used to analyze customer behavior and determine the value of customers to a business.

Each of the three letters represents a different factor:

- Recency: This refers to how recently a customer has made a purchase from the business. Customers who have made a purchase more recently are typically considered more valuable than those who have not made a purchase in a longer period of time.
- Frequency: This refers to how often a customer has made a purchase from the business. Customers who make purchases more frequently are typically considered more valuable than those who make purchases less frequently.
- Monetary Value: This refers to how much a customer has spent on purchases from the business. Customers who have spent more money are typically considered more valuable than those who have spent less.

By analyzing these three factors, businesses can segment their customer base and develop targeted marketing strategies to engage with each segment in the most effective way possible.

1.2. How RFM applied in this case:

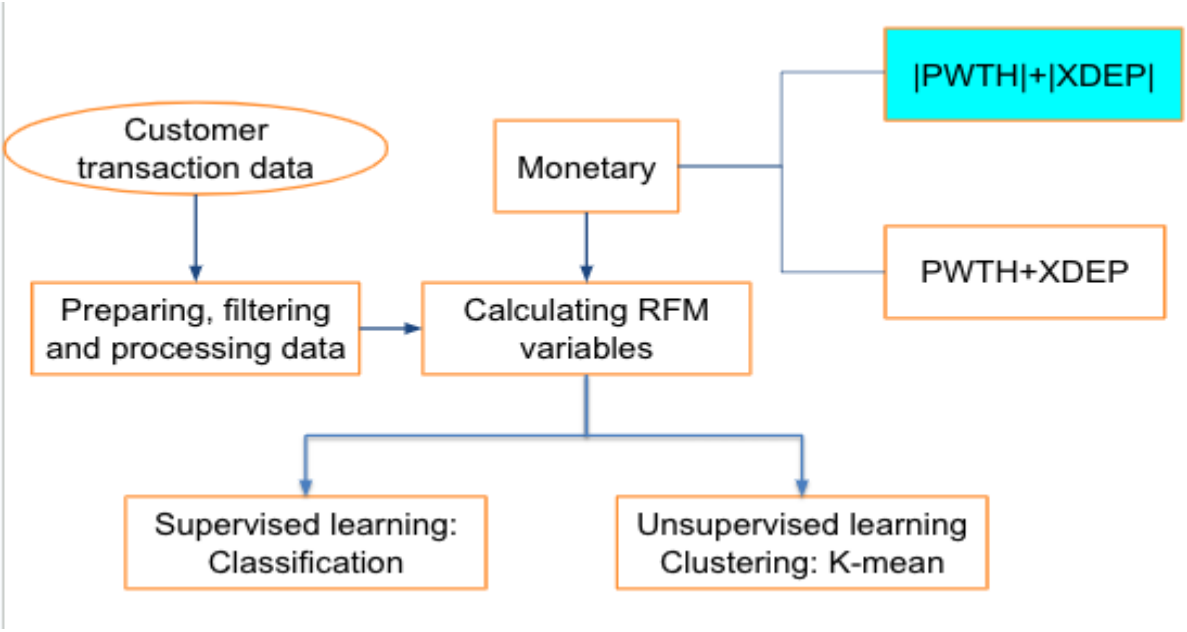


Figure 2: Project framework

Recency value is calculated by the numbers of day from the last transaction until the current date

Frequency value is the number of transactions during the time period

Monetary value is calculated by 2 different ways:

- Sum of absolute value of spending and income: This approach comes from a fee perspective with an assumption that the bank will get benefit from both type of transaction
- Sum of net value of spending and income: this approach comes from the interest perspective. If customers spend all the money they receive. The monetary value of that customer will be low.

2. RFM segmentation:

1.1. Supervised model - RFM score:

We split each factor into 4 quartiles based on their value. The RFM score for each customer will include 3 digits representing each factor.

- Recency: the higher value is, the lower score it will get

- Frequency: the higher value is, the higher score will be
- Monetary: the higher value is, the higher score will be

After that, we have 5 clusters:

Cluster	Score	Percentage	Description
Core customers	111	15%	Use their cards recently and frequently with high amount
Loyal customers	1xx	33%	use their card recently
Big Spenders	xx1	37%	big value in each transaction
At risk	134	1%	R is good, F and M is low
Lost customers	344 444	14%	Last spending long ago, few spending and little amount

1.2 Clustering- K-means:

1.2.1 What is K-means clustering:

The following stages will help us understand how the K-Means clustering technique works:

- Step 1: First, we need to provide the number of clusters, K, that need to be generated by this algorithm.
- Step 2: Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
- Step 3: The cluster centroids will now be computed.
- Step 4: Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.

  4.1 The sum of squared distances between data points and centroids would be calculated first.

  4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).

  4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.

K-means implements the Expectation-Maximization strategy to solve the problem. The Expectation-step is used to assign data points to the nearest cluster, and the Maximization-step is used to compute the centroid of each cluster.

1.2.2 Application of K-mean in the dataset:

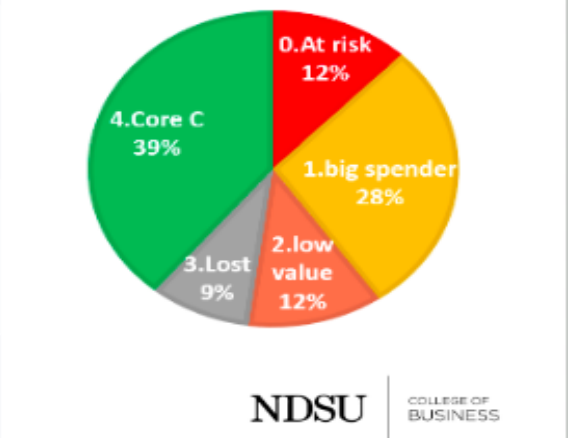
We use 3 different method to calculate the optimal value of K value and test the result with K = {3, 4, 5}:

- Elbow method
- Flattening three-dimensional graphs
- Snake plot

We come up with K = 5

Segment	Description
Core customers	R low, high F and M
Big Spender	big value in income/spending
Low value	Low F and M, but R is also low
At risk	Used to a good customers with high transaction amount and frequency, but haven't used for a while
Lost customers	Last spending long ago, few spending and little amount

cluster	R	F	M	count	per
0	421	406	45,662	21,379	12%
1	24	231	79,568	49,494	28%
2	37	42	4,316	20,839	12%
3	489	13	1,538	15,637	9%
4	19	1,601	190,317	68,525	39%



1.2.3 RFM Model using Net Transaction Amount:

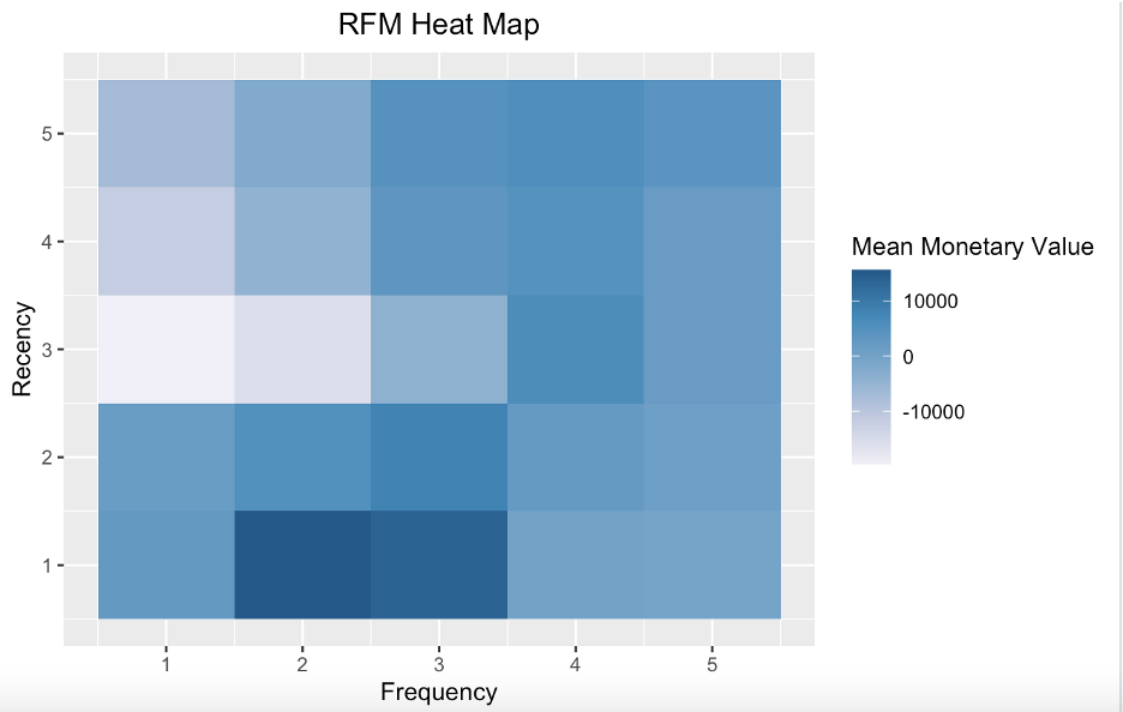
Gate City can earn interest through customer deposits as well so we used the net transaction amount to carry out RFM analysis using R-studio. The net transfer amount provides insight on how much money is coming into the bank after considering the payments. This would be helpful to consider because Banks benefit from deposits customers make in them. They can deploy those funds in the money market and other assets to earn interest and generate income.

This analysis includes all types of accounts (checking + saving + mortgage).

Using the transaction table, we carried out the RFM model. We looked at the RTXN_AMT, Date and frequency of transactions of each account and gave rfm scores to each customer using R studio. The scores range from 1-5 with 1 being the lowest and 5 being the highest.

Using the rfm model through R studio, we find that there are lots of customers that have low recency and average frequency but have a high monetary value. These customers have

the tendency to spend more. But low recency shows that these customers might be gradually stopping doing further transactions with gate city.



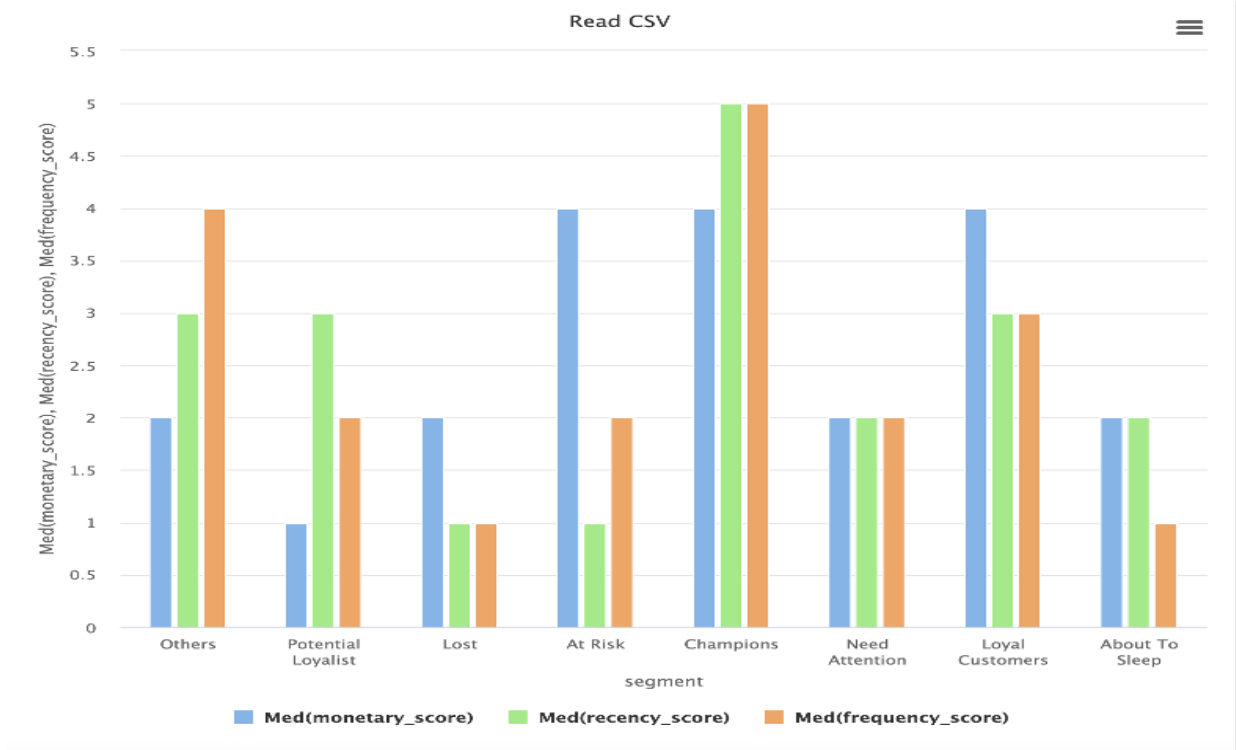
These kinds of customers, we have categorized them as “At Risk”. The company needs to focus on these customers and try to engage them again. We also see that people with low frequency and average recency have low monetary value.

1.2.4 Segmentation:

We categorized the customers into different groups regarding their RFM score in this way:

Segment	Description	R	F	M
Champions	Bought recently, buy often and spend the most	4 - 5	4 - 5	4 - 5
Loyal Customers	Spend good money. Responsive to promotions	2 - 5	3 - 5	3 - 5
Potential Loyalist	Recent customers, spent good amount, bought more than once	3 - 5	1 - 3	1 - 3
New Customers	Bought more recently, but not often	4 - 5	<= 1	<= 1
Promising	Recent shoppers, but haven't spent much	3 - 4	<= 1	<= 1
Need Attention	Above average recency, frequency & monetary values	2 - 3	2 - 3	2 - 3
About To Sleep	Below average recency, frequency & monetary values	2 - 3	<= 2	<= 2
At Risk	Spent big money, purchased often but long time ago	<= 2	2 - 5	2 - 5
Can't Lose Them	Made big purchases and often, but long time ago	<= 1	4 - 5	4 - 5
Hibernating	Low spenders, low frequency, purchased long time ago	1 - 2	1 - 2	1 - 2
Lost	Lowest recency, frequency & monetary scores	<= 2	<= 2	<= 2

We further created charts to see the different scores for recency, frequency and monetary value. These scores are all median scores.



The “at risk” customers should be the first priority to be focused on, as they have high monetary scores but low frequency and recency scores. These are some methods Gate City can apply to engage with these customers and ultimately improve their recency and frequency scores.

1. Loyalty programs: Offer loyalty programs or reward systems that encourage customers to continue using your services. You can create a points-based system or provide exclusive benefits and discounts for high-value customers.

2. Proactive outreach: Identify high-value customers who have not used your services in a while and reach out to them proactively. Offer personalized assistance, such as a dedicated account manager or a customized financial plan to help them achieve their goals.

3. Innovative services: Offer innovative services that provide value to your high-value customers. For example, you can offer investment advisory services, financial planning tools or customized credit solutions that cater to their specific needs.

4. Partnership with other businesses: Partner with other businesses that serve the same customer base to offer joint promotions or discounts. This can increase customer engagement and loyalty to both businesses.

Next, Gate City needs to focus on “potential loyalists”. Gate City needs to figure out how can they change these customers from “potential loyalists” to “loyal customers”. Also “Others” category has lots of people, that category also needs to be focused on to improve their monetary score. Here are some strategies they can apply to increase monetary score of customers in those categories:

1. Premium Accounts: Gate City can offer premium accounts to customers with low-value transactions as a way to increase their value. For example, you can offer higher interest rates, waived fees, and other exclusive benefits that can entice customers to upgrade their accounts. And the account will only be upgraded after you have spent a certain amount on your debit card in that year.

2. Education: Some customers may be hesitant to use their debit cards frequently due to concerns about security or budgeting. By providing educational resources on topics like fraud prevention and personal finance management, you can help customers feel more comfortable and confident using their debit cards.

V. Limitations:

The RFM model assumes that customers will continue to behave in the same way as they have in the past. However, customer behavior can change over time, and this can affect the accuracy of RFM scores. Therefore, Gate City should periodically evaluate and update their RFM model to ensure that it remains relevant.

Additionally, the weight given to each component of the RFM model (Recency, Frequency, Monetary) is often assumed to be equal, but this may not always be the case. Experts recommend evaluating the relative importance of each component for a particular business or industry. In case of a bank, frequency score could be of higher importance as compared to others

Furthermore, while the RFM model is useful for customer segmentation, it has limitations in providing a comprehensive understanding of customer behavior. The RFM model does not capture other important factors that influence customer behavior, such as customer demographics, customer satisfaction, and customer loyalty. Therefore, Gate City should use RFM scores in conjunction with other methods and tools to gain a more complete understanding of their customers.

Finally, the RFM model does not take into account external factors that may influence customer behavior, such as changes in the economy, competition, or seasonality. Therefore, businesses should be aware of these external factors and use additional analysis to evaluate their impact on customer behavior. Proposed different approaches and methods can be applied to the current dataset to overcome these limitations and gain deeper insights into customer behavior.

VI. Lessons Learned:

Working with Big Data (over 50 GB) can be challenging but also rewarding if approached with the right tools and frameworks. Here are some of the lessons we learned while working on such data:

PySpark: PySpark is a powerful open-source tool that allows for distributed processing of Big Data. It provides an easy-to-use interface for data processing. In this analysis, PySpark was used to clean the data as well as carry out the initial RFM framework.

R-Studio: R-Studio is a popular Integrated Development Environment (IDE) for R programming language. R-Studio was used to carry out the RFM Analysis in which we took the average transaction amount into account.

RFM Framework: The RFM (Recency, Frequency, Monetary) Framework is a popular method for customer segmentation based on historical purchase data. When working with Big Data, RFM scores were calculated using PySpark and R-Studio. However, it should be noted that the RFM framework has limitations and may not capture all important factors that influence customer behavior.

Data Visualization: Data visualization is an essential part of analyzing Big Data. It helps in understanding patterns and trends in the data, identifying outliers, and communicating insights. We used Power Bi for most of our data visualization.

VII. Conclusion:

The RFM (Recency, Frequency, Monetary) model is a popular customer segmentation method that uses historical purchase data to identify patterns in customer behavior. By analyzing each customer's recency, frequency, and monetary value of their purchases, businesses can segment their customers into different groups and develop targeted marketing strategies.

In the case of Gate City, both of the RFM models identified the "high-risk" customers that the company needs to focus on. These customers may have exhibited behaviors such as infrequent purchases, high monetary value, and low recency. By identifying these customers, the company can take proactive measures to retain them, such as personalized marketing campaigns, incentives, or special promotions.

Similarly, the high-value customers identified by the RFM models can be targeted for special promotions or loyalty programs. These customers may have exhibited behaviors such as frequent purchases, high monetary value, or long-term loyalty to the company. By offering targeted incentives and promotions to these customers, Gate City can build a stronger relationship with them and encourage repeat purchases.

Additionally, Gate City can run A/B tests on different segments of customers to see which promotions or incentives lead to the highest revenue. By testing different strategies on different segments of customers, the company can identify the most effective marketing tactics and allocate its resources accordingly. This can help the company to optimize its marketing campaigns and increase revenue.

In conclusion, the RFM model is a powerful tool for customer segmentation and can help businesses to identify high-risk and high-value customers. By developing targeted marketing strategies and running A/B tests, businesses can retain their customers, increase loyalty, and optimize their revenue.

Appendix

1. Dataset description

Customers:
27 files named NDSU_customer_1 - 27

No.	Variable	Data Type	Description	Notes
1	**Customer**	String	This is the customer ID	
2	Adddate	Date	This is the date the account was created	
3	eom_prod_dt	Date		
4	AGE	Numerical	Age of the customer	From 1-114
5	deceased	boolean	Is the customer alive	
6	household_key	String	ID of the household the customer belongs to	
7	customer_type	Categorical	Is the customer business or personal	PERSONAL BUSINESS
8	cust_status	Categorical	Is the customer still active or inactive	ACTIVE CLOSE

Account table:
Through 12/31/2020: 34 files named NDSU_Account_1 - 34

After 12/31/2020: 22 files named NDSU_aAccount_1 - 22

No.	Field names	Data type	Description	Notes
1	**Account_key**	String	Account number	Not unique
2	Active_date	Date	The date only available when we have the value in column orig_bal_amount (6)	
3	Account_status	Categorical	Status of an account	ACT APPR CLS CO DORM IACT NPFM ORIG
4	closed_date	Date	The date closed an account	
5	eom_prod_dt	Date	The end of month	
6	balance	number	The balance at the end of month	
7	orig_bal_amount	number	The initial balance deposited to the account	
8	Major_type	Categorical	Type of an account	CK CML CNS MTG SAV TD (CERTIFICATE)
9	major_description	Categorical		CERTIFICATE CHECKING COMMERCIAL LOAN CONSUMER LOAN MORTGAGE LOAN SAVINGS
10	minor_description	Text	Description of the account	
11	customer_type	Categorical	Type of customer	PERSONAL BUSINESS

12	**Household_key**	String	Household ID	Not unique

Transaction:

No.	Variable	Data Type	Description	Notes
1	ACCTKEY	String	Account number	Length is different
2	RTXNNBR	Numerical	Unique number for a transaction in each account	
3	RTXN_AMT	Numerical	Transaction amount	
4	BAL_AFTER_RTXN_AMT	Numerical	Balance after transaction amount	
5	RTXN_CNT		1: transaction clear -1: unclear/rejected/ error transaction	
6	RTXN_TYPE	Categorical	Transaction Type	
7	RTXN_TYPE_CD	Categorical	Transaction Type Code	
8	MAJOR_TYPE	Categorical	Type of an account	More types than the previous table
9	MINOR_TYP_CD	Categorical		
10	ACT_DATE_TIME	Date/Time	The time of a transaction happens	
11	EFF_DATE	Date	The date that a transaction actual effect to the balance	
12	SIGNED_BASED_TXN_FLG	Categorical	How to verify a transaction?	Yes/No
13	BRANCH	String	Branch key	
14	TXT	Text	Description of the transaction	We can extract the merchant's name of that spending
15	PARENT_ACCT_KEY	String	Where does the transfer come from?	

16	PARENT_TRAN_NBR	Numerical	Help to know with transaction kick off the transaction,	
17	RECURRING_TRAN_YN	Boolean	Is this transaction recurring?	
18	SICSUBCD	Numerical	A code for a merchant	Can use the code to group the spending, or know what purpose of the transaction is

2. Script:

2.1 Script for downloading and merging dataset from Dropbox to Google Drive

2.2 Script to calculate RFM Score and classification

2.2.1. R_script

2.2.2. PySpark

3. Visualization