Final Marketing Report on

E-commerce Multichannel Direct Messaging

Prepared by:

Group 6 (BlackPink)

120020362 Catherine Febriani

121090709 Yang Yanzhi

121020353 Fu Xiaohan

120090794 Yuwen Hu

MKT3310 Marketing Analytics

Prof. Zhang Qiang

6 March 2024

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Executive Summary

Overview

Our dataset is from REDES46, a multinational marketing intelligence and technology service provider, which has compiled data for an anonymous mid-sized Russian retail company under the project "CDP."

The Problem

Primarily, this paper aims to solve the marketing problem of optimizing the campaign's effect on consumers by adjusting the characteristics of the campaign message.

The Solution

Based on our findings, this paper suggests using email as it emerges as a vital channel for engaging customers. The email should have increased message count and detailed subjects as it shows a higher likelihood of drawing customer attention and action. It is also important to note that the segmentation of customers, based on their interaction with campaign elements, allows for more tailored marketing strategies. For example, shorter, emoji-rich subjects tend to resonate well, likely due to their ability to capture attention in a cluttered digital space.

Highlights

One interesting finding from this research was the negative correlation of over-personalization with click rates. This hints at potential consumer privacy concerns or message relevancy issues.

Table of Contents

1. Project Background

In the early 2000s, the drive towards digitized marketing led advertisers to invest heavily in online direct marketing. This era began with targeted ads across multiple media channels, such as Twitter and Facebook, and as technology evolved, this ecosystem expanded into a complex and customizable marketing method. This evolution challenges marketers to rapidly and accurately develop creative marketing campaigns. Frequently launching new campaigns necessitates thoroughly re-evaluating previous campaigns to gauge their success and adapt to new marketing dynamics. Additionally, marketers must now consider shifting consumer behaviors, such as the growing emphasis on ethical marketing practices and personalized information. In response, our group aims to analyze a retail company's campaign messages sent to users, trying to determine the most effective promotional messages encouraging online shopping behavior based on historical data. We utilized datasets referenced from Kechinov, M. (2023a, June) and Kechinov, M. (2023b, June).

2. Description of Marketing Problem and Dataset

2.1. Marketing Problem

Primarily, this paper aims to solve the marketing problem: **How to optimize the campaign's effect on consumers by adjusting the characteristics of the campaign message**. To achieve this, our group seeks to explore the relationship between users' different behaviors and the effects of varying campaign messages' characteristics on the receiver's behaviors. We also wanted to find out how to adjust the message's attributes so that the company could optimize the effect brought by the campaign message, i.e., maximize the number of users who conduct positive actions (open/ click/ purchase) toward the campaign message.

2.2. Dataset

Our dataset is from REDES46, a multinational marketing intelligence and technology service provider, which has compiled data for an anonymous mid-sized Russian retail company under the project "CDP." The data includes two datasets, the primary data resources of our analysis: **"campaigns"** and **"messages-demo."**

The "campaigns" dataset includes 19 columns and describes the characteristics of messages related to each campaign. For example, each message consists of the campaign

message's type (e.g., Bulk, Trigger, Transactional) or the subjects (e.g., emoji, bonus, etc.). This dataset attains most of our independent variables. Some columns were dropped during the analysis because we won't use them as the analysis variables. Meanwhile, the "messages-demo" includes 32 columns and depicts individual messages each user receives, detailing the messages' statuses and additional metadata. We extracted user's behaviors toward a specific message from this dataset. For example, whether the user opens/clicks on the link included in the message/purchases through the link/blocks the message... Those user's actions towards the messages were used as the dependent variables.

This project encompasses two years of multichannel messaging campaign data, specifically focusing on bulk campaigns as outlined in the "campaigns" dataset. We will utilize the "campaigns" and "messages-demo" datasets to address our marketing problems proposed in the following part. The detailed column information of "campaigns" and "messages-demo" datasets are described in Table 1 and Table 2 in the Appendix, respectively.

3. Data Description

3.1. Variable Definition, Decision Variables, and Outcome/Output Variables

The "campaigns" dataset, detailed in Appendix ***Table 1***, includes information on each campaign's type, channel, timing, and the message subject's characteristics, such as length, personalization, and thematic elements (e.g., sales or events). These elements are considered decision variables that potentially affect the outcomes of the marketing campaigns.

The "message-demo" dataset, shown in Appendix ***Table 2***, provides extensive details on user interactions with the campaigns, highlighting behaviors such as message opening, clicking, and purchasing. These behaviors are our outcome variables and are critical for assessing the effectiveness of each campaign.

3.2. Data Cleaning Process

For the “campaign” dataset (stored in df1 seen in ***Graph 1***), our primary focus was on retaining only campaigns of the "bulk" type. The dataset includes two columns, "ab_test" and "is_test," which indicate whether a campaign is for testing purposes. To align our dataset to analyze promotional campaigns, we removed rows where either of these columns had a

"True" value. Additionally, we eliminated columns irrelevant to our analysis ('warmup_mode,' 'hour_limit,' 'ab_test,' 'is_test,' 'position,' 'channel') to streamline the dataset.

The "message-demo" dataset contains more than 6 million entries of users' interactions with each campaign, which was too large to proceed. We randomly selected 20% of the data from the original dataset to optimize the data analysis process. The remaining around 1.2 million data is still enough for analysis.

Afterward, our group merged the characteristics of campaigns with user behaviors into a single dataset (df) to facilitate analysis (***Graph 2***). Upon reviewing the independent variables (IVs) through the value_counts function, we identified and removed IVs with uniform values across the dataset as they hold no analytical value ('subject_with_emoji,' 'subject_with_bonuses,' 'subject_with_saleout,' 'message_type,' 'stream'). Additionally, we converted the "channel" column into a binary variable named "is_email," where "email" was coded as one and "mobile_push" as 0, recognizing its potential impact on user behavior.

Given that most variables were categorical, represented by "t" and "f" values, we transformed them into binary variables (1 or 0) to facilitate smoother analysis.

4. Decomposition of the Marketing Problem

The marketing problem we want to solve is how to optimize the effectiveness of campaign messages to influence consumer behaviors positively, which include actions like opening, clicking, and purchasing through campaign messages. To address this, the decomposition involves breaking down the problem into analyzable components as outlined below:

1. **Consumer actions (DVs):**
    a. **Positive actions:** open, click, purchase
    b. **Negative actions:** unsubscribe, block, complain
2. **Campaign message characteristics (IVs):**
    a. **Message type:** bulk, trigger, transactional
    b. **Content features:** subject length, use of personalization, use of discounts, use of deadlines
    c. **Delivery channel (is_email=1/0):** email vs. mobile-push

4.1. Analytical Approaches

To have an in-depth understand of such problem, our group have taken two main analytical approaches**: (1) Data Description,** and **(2) Regression Analysis**

In the data description section, we aim to establish basic patterns and insights from the data, assisting with the data cleaning process to select variables for the formal analysis. Rather than using typical descriptive analytic methods discussed in class, we begin by employing the describe() function to obtain summary statistics that reveal general trends and characteristics of the data. Then, we explored the relationship between IVs and DVs by computing the mean values of IVs under each DV. After identifying the relationship between each DV, we used two graphs to visualize the pattern, which helped us gain basic insight into this dataset. corr() function was used to identify correlations between different variables, which can highlight relationships or multicollinearity issues among the IVs.

The second approach employed was logistic regression, given the binary nature of our DVs (e.g., is_opened, is_clicked, is_purchased). This method is suitable for modeling the probability of occurrence of these binary events. Our objective is to determine the effects of IVs on each DV, providing the company with recommendations on message characteristics that should be emphasized to elicit more positive feedback from users.

4.2. Types of Data Used

Two types of data were utilized throughout our research. These are:

1. **Structured data:** campaign and user interaction data stored in tables "campaigns" and "messages-demo".

2. **Binary variables:** all of the user's action variables are binary (e.g., has the message been opened? Clicked? Purchased?), making logistic regression an appropriate analytical method.

By decomposing the marketing problem as described, we can systematically approach each component with the most suitable analytical techniques, thereby enhancing our understanding and ability to make informed decisions to optimize campaign effectiveness.

5. Research Findings

5.1. Data Descriptive Analytics

Our descriptive analytics aims to understand our data's basic patterns and insights, setting a solid foundation for more effective analysis. We utilized df.describe() and df.corr() functions to examine data patterns and correlations among variables. The correlation matrix (***Graph 7***) revealed a high correlation between two IVs ("subject_with_personalization" and

“subject_with_discount”), highlighting the potential issue of multicollinearity in regression analysis. Furthermore, by comparing the mean values of IVs across different DVs (receiver's actions toward messages with distinct characteristics), we discovered fundamental trends in users’ behavior: campaigns received through email tend to have higher open and click rates. Interestingly, negative actions such as unsubscribing, blocking, complaining, and purchasing were also predominantly observed in the email channel, possibly because these actions are more feasible or visible in email interfaces than in mobile pushes. Additionally, campaigns featuring personalization and discount information were more likely to be opened and clicked, and these resulted in purchases, offering initial insights into user preferences and behavior patterns.

Observing the relationship between our DVs, we found a conversion chain between the receivers' positive responses. Users need to purchase the products after clicking on the message, and the click action should go after the open action. We created a funnel chart using the graphviz package to visualize this conversion chain. The graph shows the loss of users from opening the message to the final purchase action. Based on ***Graph 3*** in the appendix, the user rate dramatically drops in each step where “click” to “purchase” received the most percentage drop. In addition, we also found that most people who send negative feedback do not open the message, and there is no clear relationship between negative actions (unsubscribe, complain, block) of consumers. The overall relationship between DVs is shown in ***Graph 4***.

5.2. Message Segmentation

We tried to conduct a segmentation analysis to identify the patterns of messages sent through different channels ("is_email"=1: email channel; "is_email"=0: mobile-push channel in ***Graph 6***) and gain some essential insights. The segmentation analysis was conducted by comparing the mean values of the independent variables under different values of the "is_email" variable. We did not utilize the typical cluster analysis for segmentation because our dataset consists of mostly dummy variables that are better suited for cluster analysis. Furthermore, we cannot differentiate the variables used for profiling or segmenting. Thus, we considered checking the mean values of IVs for essential insights.

As we mentioned above, the positive actions of receivers showcase an apparent conversion-chain relationship (summarized in ***Table 4***). It should be more reasonable to test the click action among all the users who have opened the message (is_opened=1) and test the purchase action among all clicked users (is_clicked=1). The open action is checked among

the original dataset, including all receivers. Based on this analysis logic, we checked the message segments under three different data groups: 1) the original dataset; 2) the dataset contains only is_opened=1; 3) the dataset contains only is_clicked=1.

5.3. Logistic Regression

In this part, we leveraged logistic regression to explore the relationships between the characteristics of messages and user behaviors. This choice of the regression model was motivated by binary DVs within our dataset, making logistic regression an ideal analytical tool for this context. Our objective was to uncover insights into how specific message attributes influence user actions—including positive actions (open, click, purchase) and negative actions (unsubscribe, block, complain) so that we can address the challenge of enhancing the effectiveness of bulk message marketing.

Through the previous analysis, we have cleaned the data. We first wanted to check the patterns between variables through a correlation matrix to determine the IVs and DVs that could be used in the regression analysis. The result of the correlation matrix shown in ***Graph 7*** shows that "subject_with_personalization" and "subject_with_discount" are highly correlated and can lead to multicollinearity issues during regression analyses. Therefore, we deducted subject_with_discount" during the logistic regression process . In addition, the variable "total_count" describing the total number of receivers of each campaign message is meaningless in the real world since the receiver has no idea about the total count. Thus, we decided to deduct this variable from IVs as it should not affect the actions of users.

Next, our group analyzed open action among all the messages sent to users, the effect of IVs on the click action among all the opened messages, and purchase action among all the clicked messages. These steps addressed the conversion chain between three positive actions of users. Afterward, all three negative-feedback DVs were regressed across three sample sets except "is_blocked," which only ran a regression on the first two,b he was because it has no intersection with the clicked-message dataset.. finally, our group analyzed purchase action. Here, we found that the variable "subject_with_deadline" contains only identical values in this clicked-message dataset, so we selected this variable when we did the regression analysis for purchase action.

In summary, the DVs used in the logistic regression analysis include " is_email," "subject_length," subject_with_personalization", "and subject_with_deadline" (with the last variable omitted during the purchase action) analysis . Our IVs are all the user's actions

toward the message which are:"is_opened", "is_clicked", "is_purchased", is_unsubscribed", "is_complained", and"is_blocked",

5.3.1. Logistic Regression Results:

The final logistic regression results are shown in ***Graph 8*** while the discussion of the positive feedbacks' results are as follows:

DV "is_opened", three IVs: " subject_length", "is_email" and "subject_with_deadline" show a significantly positive effect on "is_opened". **For DV "is_clicked"**, two IVs: " subject_length" and "is_email" show a significantly positive effect on "is_clicked". "subject_with_personalization" and "is_clicked" are negatively correlated and sig the effects arenificant. **For DV is "is_purchased"**, only one feature, "subject_length", have ssnificant positive effect on user's purchasing behavior. The telationships between the remaining characteristics and purchasing behavior are not statistically significant.

Here are the other three negative feedback's' resuts. **For DV "is_unsubscribed"**, the result shows that in the entire sample set "subject_length", "is_email" and "subject_with_deadline" have a significant negative correlation with "is_unsubscribed" whereas only "subject_length" has a significant negative correlation with the DV in the open dated-messageaset. **For DV "is_complained" and DV "is_blocked"**, the p-values for all indeIx Vs greater than 0.05, indicating that the effects of these indeIx Vs not statistically significant. So there is o significant relationship between IVs and user's complaining and blocking behavior. (***Table 5***)

5.3.2. Interpretation of Results and Marketing Insights

Our group can interpret five critical points with marketing insights based on the findings. First, our analysis confirms that emails are an effective channel for marketing as they significantly increase the likelihood of users opening and clicking through messages. This preference suggests that emails are a familiar and trusted medium for receiving marketing communications. Second, sending more emails and extending their length can enhance user engagement, including opens and clicks. However, an excessive volume of messages may lead to a decreased purchase rate and an increased unsubscribe rate. Third, incorporating deadlines within messages was effective in encouraging users to open messages and, to some extent, preventing unsubscribes. Deadlines create a sense of urgency, prompting quicker user engagement. Fourth, a surprising finding from our analysis indicated that highly personalized messages might deter clicks. This could be due

to users' concerns over privacy and information security, highlighting the importance of balancing personalization with user comfort. Finally, our studies suggest that overly lengthy messages may overwhelm or frustrate users, leading to negative feedback. While we did not find a direct link between message characteristics and user complaints or blocks, the data revealed a positive correlation between message length and complaints.

6. Conclusions and Recommendations

6.1. Business Recommendations:

Our analysis underscores the effectiveness of email as a critical marketing channel, suggesting that businesses should focus on email to distribute their marketing content. By carefully adjusting the volume and length of these messages, companies can enhance user engagement while reducing the risk of adverse reactions. However, it's essential to balance the quantity and quality of information. While providing more information can captivate and convert potential users, overloading them may lead to disengagement. Therefore, marketers are encouraged to craft messages that are both concise and rich in content, ensuring they are relevant and valuable to the audience. Personalization in messaging has received mixed responses, indicating the need for a careful, respectful approach to customization that does not infringe on user privacy or lead to discomfort.

Additionally, the association of longer messages with an increased rate of complaints suggests a demand for clarity and succinctness in communication. By ensuring that messages are straightforward and focused, companies can mitigate the chances of user complaints, fostering a more positive engagement with their audience. This holistic approach to email marketing, emphasizing a thoughtful balance of volume, content quality, and personalization, can significantly improve marketing outcomes.

6.2. Main Findings and Addressing The Marketing Problems:

Our data analysis, cluster analysis, logistic regression, and classification methods reveal intricate customer behavior patterns in response to direct marketing campaigns. The channel of communication and message characteristics—like subject length, use of emojis, and personalization—have varied impacts on customer actions, such as opening messages, clicking through, and making purchases more visible with bulk campaigns before holidays.

Otherwise, our group concludes that email emerges as a vital channel for engaging customers, with increased message count and detailed subjects showing a higher likelihood of

drawing customer attention and action. An interesting counterpoint is the negative correlation of over-personalization with click rates, hinting at potential consumer privacy concerns or message relevancy issues. On the other hand, the segmentation of customers based on their interaction with campaign elements allows for more tailored marketing strategies. For example, shorter, emoji-rich subjects tend to resonate well, likely due to their ability to capture attention in a cluttered digital space.

In conclusion, our analysis has addressed the marketing problem by providing a roadmap to align with customer preferences and behaviors. The analysis has directly responded to the need for more effective bulk message marketing by identifying which message characteristics bolster engagement and purchase behavior. While the correlation between message characteristics and adverse outcomes like complaints or unsubscriptions was less evident, the positive associations offer clear directives for optimizing message content.

6.3. Limitations and Recommendations for Future Research:

Some critical limitations of this research should be addressed to better future research on the same subject. First, our group discovered a problem with the dataset's small number of IVs. This may have resulted in a low fit with our logistic regression models, thus only explaining a small part of the variation. To counter such a problem, it is recommended to leverage advanced machine learning techniques to predict behaviors with greater accuracy in experiments.

Second, our group recommends further research on the reason for the negative correlation of over-personalization with click rates. This can be done by more granular A/B testing to refine our understanding of how specific message characteristics influence customer behavior across different segments. Furthermore, this finding may have produced a new marketing problem of understanding how consumers perceive "personalization" and how marketers should acceptably approach this matter. Third, future studies should note our limited results from the negative feedback, perhaps resulting from limited relevant data. One way to improve this approach is by collecting more data via longitudinal studies to observe changes in consumer behavior over time. This may offer insights into the evolving effectiveness of marketing strategies.

References

Kechinov, M. (2023a, June). Direct messaging campaigns dataset overview. Kaggle.com. https://www.kaggle.com/code/mkechinov/direct-messaging-campaigns-dataset-overview/notebook#Dataset-overview

Kechinov, M. (2023b, June). E-commerce multichannel direct messaging 2021-2023. Kaggle.com https://www.kaggle.com/datasets/mkechinov/direct-messaging/data?select=messages-demo.csv

Appendix

Estimation of Time Spent

No.	Activity	Time Spent
1	Cluster and Factor Analysis	7 days
2	Consultation with professor	1 day
3	Logistic Regression	7 days
4	Consultation with professor	1 day
5	Conclusions and Recommendations	5 days
6	Clean coding notebook	5 days
7	Report for Data Analysis	5 days
8	PPT Making and Practice	7 days
9	Final report	7 days
10	Clean coding notebook	5 days

Table 1: Data description for campaign dataset

Column Name	Description	Variables Type
id	Unique campaign ID only for the specific campaign type	
campaign_type	Campaign type (bulk, trigger, transactional)	
channel	Channel (email, mobile_push, web_push, sms)	decision variables
topic	Meaning of a campaign (sale out, happy birthday, etc.)	decision variables
started_at	Bulk campaign start datetime	decision variables
finished_at	Bulk campaign finish datetime	decision variables
total_count	Total recipients in bulk campaign	decision variables
subject_length	Email subject length	decision variables
subject_with_personalization	Subject contains recipient's name	decision variables
subject_with_deadline	Subject has deadline meaning	decision variables
subject_with_emoji	Subject has emoji symbols	decision variables
subject_with_bonuses	Subject mentions bonuses for actions	decision variables
subject_with_discount	Subject mentions a discount	decision variables
subject_with_saleout	Subject mentions a sale out	decision variables

Table 2: Data description for "message-demo" dataset

Column Name	Description	Variables Type
id	Message sequence ID *will not be used	
message_id	Message unique ID	
campaign_id	Campaign ID from campaigns.csv)	
message_type	Campaign type (bulk, trigger, transactional)	decision variables
client_id	Client ID	
channel	Message channel (email, web_push, mobile_push, sms)	decision variables
category	Category *will not be used	
platform	Device type used to open a message	
email_provider	Public email provider (for email messages)	
stream	Additional identifier of data source (desktop, ios and android)	
date	date in YYYY-MM-DD when a message was sent	
sent_at	Datetime when a message was sent	
is_opened	Boolean flag if a message was opened by a recipient	outcome variables
opened_first_time_at	First time when a message was opened	
opened_last_time_at	Last time when a message was opened (can be equal to opened_first_time_at, if the message was opened only once)	
is_clicked	Boolean flag if a message was clicked by a recipient	outcome variables
clicked_first_time_at	First time when a message was clicked	
clicked_last_time_at	Last time when a message was clicked (can be equal to clicked_first_time_at, if the message was clicked only once)	
is_unsubscribed	Boolean flag if a recipient clicked unsubscribe link in a message	outcome variables
unsubscribed_at	Datetime when a recipient clicked unsubscribe link in a message	
is_hard_bounced	Whether the message was hard bounced	
is_soft_bounced	Whether the message was soft bounced	
soft_bounced_at	Datetime when a message was "soft bounced"	
is_complained	Boolean flag if a recipient clicked SPAM button in email client	outcome variables
complained_at	Datetime when the message has been complained	
is_blocked	Boolean flag if a delivery attempt was temporarily blocked by email provider	outcome variables
blocked_at	Datetime when a delivery attempt was temporarily blocked by email provider	
is_purchased	Boolean flag if a recipient clicked any link in a message, opened a website or mobile app and made a purchase	outcome variables
purchased_at	Datetime when a recipient made a purchase after click on email or other message	
created_at	Datetime when the message is created *will not be used	
updated_at	Datetime when the message is updated *will not be used	

Table 3: Data Description for Cleaned Dataset

Column Name	Description	Variables Type
id	Message sequence ID *will not be used	
campaign_id	Campaign ID from campaigns.csv	
is_email	Message channel (email, web_push, mobile_push, sms)	decision variables
is_opened	Boolean flag if a message was opened by a recipient	outcome variables
is_clicked	Boolean flag if a message was clicked by a recipient	outcome variables
is_unsubscribed	Boolean flag if a recipient clicked unsubscribe link in a message	outcome variables
is_complained	Boolean flag if a recipient clicked SPAM button in email client	outcome variables
is_blocked	Boolean flag if a delivery attempt was temporarily blocked by email provider	outcome variables
is_purchased	Boolean flag if a recipient clicked any link in a message, opened a website or mobile app and made a purchase	outcome variables
total_count	Total recipients in bulk campaign	decision variables
subject_length	Email subject length	decision variables
subject_with_personalization	Subject contains recipient's name	decision variables
subject_with_deadline	Subject has deadline meaning	decision variables
subject_with_discount	Subject mentions a discount	decision variables

Table 4: Conversion chain among Open, Click, and Purchase

Stage	IV	Sample Set
Stage 1	**is_opened**	The whole sample set
Stage 2	**is_clicked**	The opened dataset (is_open = 1)
Stage 3	**is_purchased**	The click dataset (is_click = 1)

Table 5: Negative feedback results in three sample sets

DVs	The Whole Dataset	The Open Dataset	The Click Dataset
Unsubscribe	Significant negative correlation	Significant negative correlation	No significant relationships
Complained	No significant relationships	No significant relationships	No significant relationships
Block	No significant relationships	No significant relationships	No overlap

Graph 1: Data info of campaign dataset (df1)

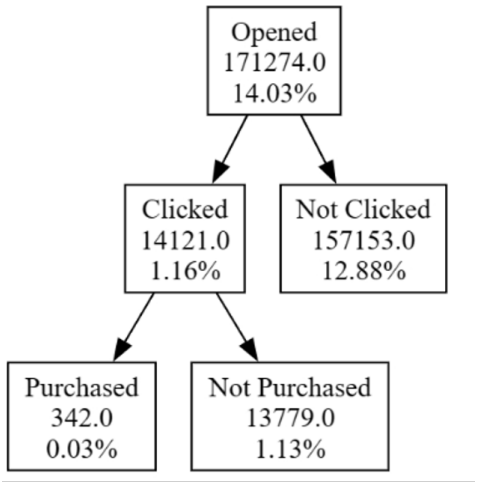
```
In [38]: df1.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1818 entries, 0 to 1829
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   campaign_id                   1818 non-null   int64
 1   campaign_type                 1818 non-null   object
 2   topic                         1798 non-null   object
 3   started_at                    1818 non-null   object
 4   finished_at                   1802 non-null   object
 5   total_count                   1818 non-null   float64
 6   subject_length                1818 non-null   float64
 7   subject_with_personalization  1818 non-null   object
 8   subject_with_deadline         1818 non-null   object
 9   subject_with_emoji            1818 non-null   object
 10  subject_with_bonuses          1818 non-null   object
 11  subject_with_discount         1818 non-null   object
 12  subject_with_saleout          1818 non-null   object
dtypes: float64(2), int64(1), object(10)
memory usage: 198.8+ KB
```

Graph 2: Data info of merging dataset (df): [merge(df2, df1)]

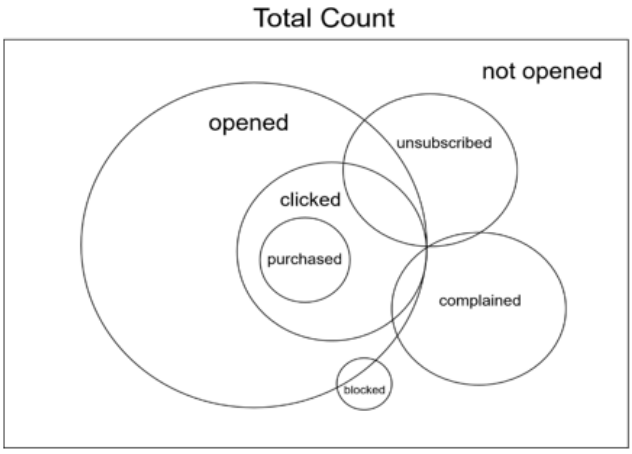
```
In [39]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 952896 entries, 0 to 952895
Data columns (total 14 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   id                           952896 non-null  float64
 1   campaign_id                  952896 non-null  float64
 2   is_email                     952896 non-null  float64
 3   is_opened                    952896 non-null  float64
 4   is_clicked                   952896 non-null  float64
 5   is_unsubscribed              952896 non-null  float64
 6   is_complained                952896 non-null  float64
 7   is_blocked                   952896 non-null  float64
 8   is_purchased                 952896 non-null  float64
 9   total_count                  952896 non-null  float64
 10  subject_length               952896 non-null  float64
 11  subject_with_personalization 952896 non-null  float64
 12  subject_with_deadline        952896 non-null  float64
 13  subject_with_discount        952896 non-null  float64
dtypes: float64(14)
memory usage: 101.8 MB
```

Graph 3: Conversion chain and conversion rate between 3 positive-action DVs



Graph 4: Relationship between all DVs



Graph 5: Relationship between IVs and 6 DVs

```
In [24]: var = ['is_email','total_count', 'subject_length', 'subject_with_personalization','subject_with_deadline', 'subject_with_discount']
         df.groupby("is_opened")[var].mean()
```

Out[24]:

is_opened	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.343609	712449.352848	107.488693	0.000827	0.007246	0.000827
1.0	0.437165	690666.824807	113.332783	0.002505	0.008472	0.002505

```
In [25]: df.groupby("is_clicked")[var].mean()
```

Out[25]:

is_clicked	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.349711	709791.005215	108.075805	0.001012	0.007505	0.001012
1.0	0.957085	678335.098081	128.218327	0.005382	0.000000	0.005382

```
In [26]: df.groupby("is_purchased")[var].mean()
```

Out[26]:

is_purchased	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.356559	709407.896373	108.303385	0.001061	0.00742	0.001061
1.0	1.000000	653827.441520	127.883041	0.005848	0.00000	0.005848

```
In [27]: df.groupby("is_unsubscribed")[var].mean()
```

Out[27]:

is_unsubscribed	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.372774	700049.163317	108.752872	0.001114	0.007712	0.001114
1.0	0.026409	901874.492460	99.161834	0.000018	0.001363	0.000018

```
In [28]: df.groupby("is_complained")[var].mean()
```

Out[28]:

is_complained	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.356548	709424.959067	108.304796	0.001063	0.00742	0.001063
1.0	1.000000	599392.618785	122.044199	0.000000	0.00000	0.000000

```
In [29]: df.groupby("is_blocked")[var].mean()
```

Out[29]:

is_blocked	is_email	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.356737	709393.679753	108.308847	0.001063	0.007418	0.001063
1.0	1.000000	377663.200000	114.400000	0.000000	0.000000	0.000000

Graph 6: Segments (is_email=1/0) in three different datasets

```
In [45]: df_nor['is_email'].value_counts()
Out[45]: 0.0    785031
         1.0    435362
         Name: is_email, dtype: int64

In [46]: df_nor.groupby('is_email').mean()
```

is_email	is_opened	is_clicked	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	0.122796	0.000772	791069.834254	100.721475	0.000000	0.011532	0.000000
1.0	0.171983	0.031043	562114.009282	121.990227	0.002979	0.000000	0.002979

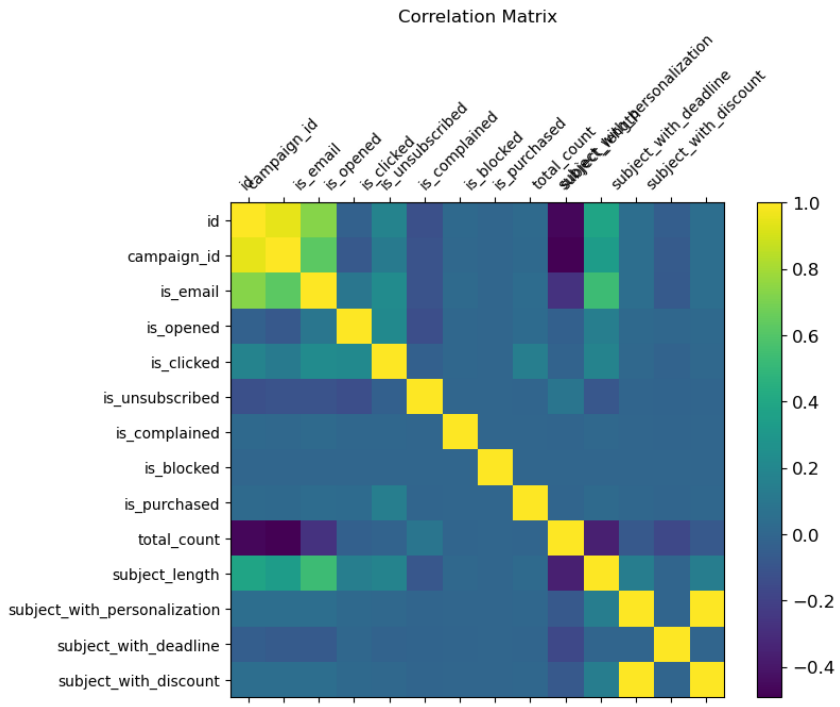
```
In [47]: df_nor.loc[df_nor.is_opened==1,'is_email'].value_counts()
Out[47]: 0.0    96399
         1.0    74875
         Name: is_email, dtype: int64

In [48]: df_nor[df_nor.is_opened==1].groupby('is_email').mean()
```

is_email	is_opened	is_clicked	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	1.0	0.006286	753415.340688	103.645577	0.00000	0.015052	0.00000
1.0	1.0	0.180501	609880.258097	125.804728	0.00573	0.000000	0.00573

```
In [49]: df_nor.loc[df_nor.is_clicked==1,'is_email'].value_counts()
Out[49]: 1.0    13515
         0.0      606
         Name: is_email, dtype: int64

In [50]: df_nor[df_nor.is_clicked==1].groupby('is_email').mean()
```

is_email	is_opened	is_clicked	total_count	subject_length	subject_with_personalization	subject_with_deadline	subject_with_discount
0.0	1.0	1.0	105707.316832	104.169967	0.000000	0.0	0.000000
1.0	1.0	1.0	700876.676730	129.296633	0.005623	0.0	0.005623

Graph 7: Correlation matrix to identify possible multicollinearity issues



Graph 8: Logistic regression results

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              is_opened   No. Observations:              1220393
Model:                         Logit   Df Residuals:                  1220388
Method:                          MLE   Df Model:                            4
Date:                Thu, 02 May 2024   Pseudo R-squ.:                0.01283
Time:                        16:34:43   Log-Likelihood:            -4.8863e+05
converged:                       True   LL-Null:                   -4.9497e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -3.2245      0.015   -209.049      0.000      -3.255      -3.194
subject_length                   0.0122      0.000     85.105      0.000       0.012       0.013
is_email                         0.1560      0.006     26.367      0.000       0.144       0.168
subject_with_personalization     0.1030      0.060      1.721      0.085      -0.014       0.220
subject_with_deadline            0.2363      0.029      8.187      0.000       0.180       0.293
================================================================================================

                           Logit Regression Results
==============================================================================
Dep. Variable:              is_clicked   No. Observations:               171274
Model:                         Logit   Df Residuals:                   171269
Method:                          MLE   Df Model:                            4
Date:                Thu, 02 May 2024   Pseudo R-squ.:                 0.2172
Time:                        16:34:44   Log-Likelihood:                -38169.
converged:                      False   LL-Null:                       -48763.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -9.0473      0.115    -78.714      0.000      -9.273      -8.822
subject_length                   0.0358      0.001     40.632      0.000       0.034       0.038
is_email                         3.0050      0.042     71.020      0.000       2.922       3.088
subject_with_personalization    -2.1123      0.136    -15.476      0.000      -2.380      -1.845
subject_with_deadline          -11.5545    111.164     -0.104      0.917    -229.431     206.322
================================================================================================

                           Logit Regression Results
==============================================================================
Dep. Variable:            is_purchased   No. Observations:                14121
Model:                         Logit   Df Residuals:                    14117
Method:                          MLE   Df Model:                            3
Date:                Thu, 02 May 2024   Pseudo R-squ.:                 0.01237
Time:                        16:34:48   Log-Likelihood:                -1590.4
converged:                      False   LL-Null:                       -1610.3
Covariance Type:            nonrobust   LLR p-value:                 1.159e-08
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      -19.7526   1760.877     -0.011      0.991   -3471.008    3431.503
subject_length                  -0.0187      0.006     -3.179      0.001      -0.030      -0.007
is_email                        18.4981   1760.877      0.011      0.992   -3432.757    3469.753
subject_with_personalization     1.0994      0.794      1.385      0.166      -0.457       2.655
================================================================================================
```

In the whole sample set (both is_opened=0 & is_opened=1)

Logit Regression Results

Dep. Variable:	is_unsubscribed	No. Observations:	1220393
Model:	Logit	Df Residuals:	1220388
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.08868
Time:	16:34:55	Log-Likelihood:	-2.0848e+05
converged:	True	LL-Null:	-2.2877e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.0329	0.022	-91.324	0.000	-2.077	-1.989
subject_length	-0.0054	0.000	-24.600	0.000	-0.006	-0.005
is_email	-2.9767	0.027	-111.130	0.000	-3.029	-2.924
subject_with_personalization	-1.1516	1.001	-1.151	0.250	-3.113	0.810
subject_with_deadline	-2.1330	0.115	-18.619	0.000	-2.357	-1.908

In the opened sample set (only is_opened=1)

Logit Regression Results

Dep. Variable:	is_unsubscribed	No. Observations:	171274
Model:	Logit	Df Residuals:	171269
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.1438
Time:	16:34:57	Log-Likelihood:	-5340.7
converged:	False	LL-Null:	-6237.8
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-25.8318	5679.447	-0.005	0.996	-1.12e+04	1.11e+04
subject_length	-0.0291	0.003	-9.772	0.000	-0.035	-0.023
is_email	25.1632	5679.447	0.004	0.996	-1.11e+04	1.12e+04
subject_with_personalization	-0.0029	1.019	-0.003	0.998	-2.000	1.994
subject_with_deadline	0.3296	4.46e+04	7.38e-06	1.000	-8.75e+04	8.75e+04

In the clicked sample set (only is_clicked=1)

Logit Regression Results

Dep. Variable:	is_unsubscribed	No. Observations:	14121
Model:	Logit	Df Residuals:	14117
Method:	MLE	Df Model:	3
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.008381
Time:	16:34:58	Log-Likelihood:	-306.76
converged:	False	LL-Null:	-309.36
Covariance Type:	nonrobust	LLR p-value:	0.1587

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-14.0144	81.477	-0.172	0.863	-173.707	145.678
subject_length	-0.0132	0.016	-0.811	0.417	-0.045	0.019
is_email	10.0308	81.472	0.123	0.902	-149.651	169.712
subject_with_personalization	-17.5632	1.85e+04	-0.001	0.999	-3.63e+04	3.62e+04

In the whole sample set (both is_opened=0 & is_opened=1)

Logit Regression Results

Dep. Variable:	is_complained	No. Observations:	1220393
Model:	Logit	Df Residuals:	1220388
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.1134
Time:	16:35:09	Log-Likelihood:	-2928.1
converged:	False	LL-Null:	-3302.5
Covariance Type:	nonrobust	LLR p-value:	9.316e-161

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-24.9378	262.039	-0.095	0.924	-538.525	488.650
subject_length	0.0024	0.005	0.455	0.649	-0.008	0.013
is_email	17.5616	262.039	0.067	0.947	-496.025	531.148
subject_with_personalization	-11.2186	243.517	-0.046	0.963	-488.503	466.065
subject_with_deadline	-0.3935	2938.269	-0.000	1.000	-5759.295	5758.508

In the opened sample set (only is_opened=1)

Logit Regression Results

Dep. Variable:	is_complained	No. Observations:	171274
Model:	Logit	Df Residuals:	171269
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.1004
Time:	16:35:25	Log-Likelihood:	-846.14
converged:	False	LL-Null:	-940.53
Covariance Type:	nonrobust	LLR p-value:	9.702e-40

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-30.3142	1.63e+04	-0.002	0.999	-3.19e+04	3.18e+04
subject_length	-0.0057	0.009	-0.620	0.535	-0.024	0.012
is_email	24.5377	1.63e+04	0.002	0.999	-3.18e+04	3.19e+04
subject_with_personalization	-19.2481	2.22e+04	-0.001	0.999	-4.35e+04	4.35e+04
subject_with_deadline	0.0031	1.38e+05	2.26e-08	1.000	-2.7e+05	2.7e+05

In the clicked sample set (only is_clicked=1)

Logit Regression Results

Dep. Variable:	is_complained	No. Observations:	14121
Model:	Logit	Df Residuals:	14117
Method:	MLE	Df Model:	3
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.006269
Time:	16:35:26	Log-Likelihood:	-74.749
converged:	False	LL-Null:	-75.221
Covariance Type:	nonrobust	LLR p-value:	0.8150

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-22.4011	4497.744	-0.005	0.996	-8837.817	8793.014
subject_length	-0.0087	0.038	-0.232	0.817	-0.082	0.065
is_email	16.2151	4497.743	0.004	0.997	-8799.199	8831.629
subject_with_personalization	-16.1781	1.85e+04	-0.001	0.999	-3.62e+04	3.62e+04

In the whole sample set (both is_opened=0 & is_opened=1)

Logit Regression Results

Dep. Variable:	is_blocked	No. Observations:	1220393
Model:	Logit	Df Residuals:	1220388
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.1001
Time:	16:35:53	Log-Likelihood:	-60.319
converged:	False	LL-Null:	-67.026
Covariance Type:	nonrobust	LLR p-value:	0.009417

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-16.3201	115.470	-0.141	0.888	-242.636	209.996
subject_length	-0.0963	0.067	-1.437	0.151	-0.228	0.035
is_email	16.2691	115.562	0.141	0.888	-210.228	242.766
subject_with_personalization	-4.4675	1958.344	-0.002	0.998	-3842.751	3833.816
subject_with_deadline	3.6917	1108.431	0.003	0.997	-2168.792	2176.176

In the opened sample set (only is_opened=1)

Logit Regression Results

Dep. Variable:	is_blocked	No. Observations:	171274
Model:	Logit	Df Residuals:	171269
Method:	MLE	Df Model:	4
Date:	Thu, 02 May 2024	Pseudo R-squ.:	0.09847
Time:	16:35:57	Log-Likelihood:	-11.766
converged:	False	LL-Null:	-13.051
Covariance Type:	nonrobust	LLR p-value:	0.6321

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-148.1339	3.24e+04	-0.005	0.996	-6.36e+04	6.33e+04
subject_length	0.7369	83.840	0.009	0.993	-163.587	165.061
is_email	39.3595	2.94e+04	0.001	0.999	-5.76e+04	5.77e+04
subject_with_personalization	-46.4408	4402.132	-0.011	0.992	-8674.460	8581.579
subject_with_deadline	45.9363	2.97e+04	0.002	0.999	-5.82e+04	5.83e+04