Data Analysis Project Proposal

Author : Sahand Niasti

March 2020

Table of Contents

1 Introduction

1.1 Industry background

Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser. Consumers find a product of interest by visiting the website of the retailer directly or by searching among alternative vendors using a shopping search engine, which displays the same product's availability and pricing at different e-retailers. As of 2020, customers can shop online using a range of different computers and devices, including desktop computers, laptops, tablet computers, smartphones, and smart speakers.

So, E-commerce (electronic commerce) is the activity of electronically buying or selling of products on online services or over the Internet. E-commerce is in turn driven by the technological advances of the semiconductor industry, and is the largest sector of the electronics industry.

Consumer to consumer (C2C) is one of the most popular approaches in markets which provide an innovative way to allow customers to interact with each other. Traditional markets require business to customer relationships, in which a customer goes to the business in order to purchase a product or service. In customer to customer markets, the business facilitates an environment where customers can sell goods or services to each other.

1.2 Data background

A fictional website 'sell-your-stuff.com' allows clients to buy and sell items that are not being used anymore (chairs, laptops, books, etc.). Some of the clients make their living by buying products refurbishing and then reselling It. But most are only interested in selling things they don't use anymore or buying cheap stuff.

The process is quite simple, a client creates an account (choosing one of the possible account types, each account type follows specific rules and legislations for different types of products and countries). One client can have only one account. After account creation clients need to deposit money to buy/sell products, as the website itself works as a broker and guarantees that both clients will be satisfied.

Internally the company keeps information of every product bought and sold splitting them into categories and socioeconomic factors. This data is then sold to third parties following data protection legislations (no personal information is disclosed). As a rule of thumb, the company sells the data by roughly 1% of the total clients have bought. (e.g. a client that buys US$ 100.00 worth of products will generate US$ 1.00 to the company).

2 Business Problem

As the information contained on these tables can reach terabytes, summary tables were created in order to help data analysts and managers to take decisions regarding marketing strategies and resources allocation. The most used tables are:

- client: Basic anonymized information about clients and client's account.
- transactions: Summary of transactions amounts of each account aggregated daily.
- campaigns: High level description of past marketing campaigns, dates and budget.

The management of sell-your-stuff.com wants to set up marketing strategies for the next fiscal year based on the summary tables. The main goal is to analyse the performance of past campaigns, coupons, profit by countries, etc. and make a report showing information to support your insights.

3 Methodology

In this project, as we said before, the sole aim is to boost the site's performance to gain more profit. So, in order to reach this desirable goal, the data extracted from the customers need to be thoroughly analyzed. At first step, we need to find out exactly what questions are we looking for to answer, how we hypothesize about the answers and finally what is our approach to fulfill the answers. All of which will be answered in this section. In the end, an entity relationship diagram of data tables will be drawn to illustrate the dataset more clearly

3.1 Questions to answer

- How was the performance of the campaigns during the past two years?
    - There were many campaigns throughout the world but which of them can be categorized as successful and which ones are not.
- When the campaigns were run, how did the customers behave during each campaign?
    - They may ignore the campaign, or be attracted to buy and sell on the site
- Is there any relationship between campaigns' dates and clients?
    - This can show us whether the campaign had any relationship with each account number and its characteristics
- Which seasons seems to be more attractive for customers to buy or sell? Does it make any difference at all?

3.2 Hypothesis

- Not all the campaigns can be categorized as successful. There must be some campaigns which have performed under or above our expectations. Can be useful to find out which ones are.
- Based on some initial data exploration, customers have 3 different approaches. Just creating an account, the ones who make a deposit and finally the ones who buy and sell like a normal person.
- There must be some alignments between transaction dates and campaign dates, as it is the main goal of campaigns to attract more customers.
- In my opinion, before and after new year can be considered a higher site traffic. But let's see what data can tell us.
- As the data given is in a two years period, some proper information can be gained to boost our sell.

3.3 Approaches

The relationships used are between type of customers, the region and coupon indicator of each client. Total buy and sell and transaction date were the key factors in analyzing transactions. And finally relationships between start date, end date and total spent was analyzed as well.

The metric measures are: comparison of campaigns held in each country, count the number of orders each customer made, total expenditure based on each season.

3.4 ERD

An entity relationship diagram (ERD) shows the relationships of entity sets stored in a database. Here we can see how our 3 tables are related to each other and how we can link them to extract profitable information.[1]

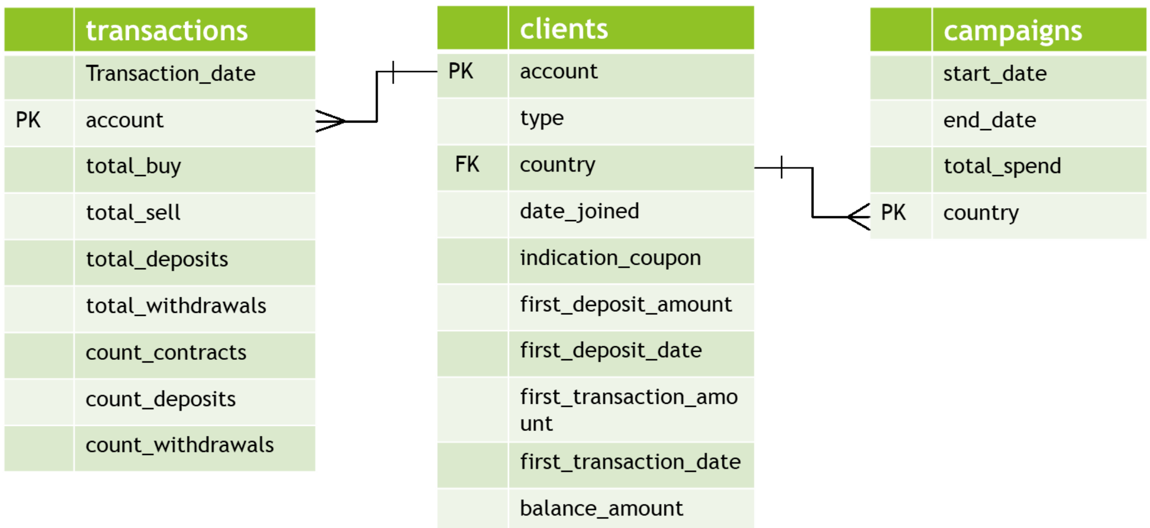

Figure 1- entity relationship diagram of dataset

4 Results

In this section, we will walk through our dataset by importing, cleaning and making it ready for further analysis. These steps will be explained concisely. Then in exploratory step, we will scrutinize each table separately to see which useful information can be extracted and help us understand the data better. After that, more advanced analytic approaches will be done to see relationships between tables, analyze the performance of each campaign, impact of each campaign on transactions and clients and finally clustering customers based on information gained in previous sections.

4.1 Importing Dataset

Three different tables related to our site required to be imported in order to be ready for further analysis. each table has been concisely introduced below:

- clients: Basic anonymized information about clients and client's account.
- transactions: Summary of transactions amounts of each account aggregated daily.

[1] Each column has been explained separately on Jupyter Notebook related to the project

- campaigns: High level description of past marketing campaigns, dates and budget

The data was primarily a SQLite3 database. Each table was exported separately to examine closely by python programming language in a Jupyter notebook. after importing the dataset as a pandas DataFrame, basic profiling data was extracted from each table to understand data a bit better.

```
In [2]: clients = pd.read_csv('clients.csv')
        clients.head()
```

Out[2]:

	account	type	residence	date_joined	indication_coupon	first_deposit_amount	first_deposit_date	first_transaction_amount	first_transaction_date
0	5019	type_6	ua	2018-06-18	308	9090.91	2018-06-18	NaN	NaN
1	3655	type_6	ve	2017-09-04	83	9090.91	2017-09-04	0.120000	2017-09-04
2	2809	type_6	br	2017-01-16	308	9090.91	2017-01-16	0.205556	2017-01-16
3	3876	type_6	mm	2017-11-01	232	9090.91	2017-11-01	NaN	NaN
4	4973	type_6	lk	2018-06-09	308	9090.91	2018-06-09	0.110000	2018-06-11

Table 1 - clients table

```
In [4]: transactions = pd.read_csv('transactions.csv')
        transactions.head()
```

Out[4]:

	transaction_date	account	total_buy	total_sell	total_deposits	total_withdrawals	count_contracts	count_deposits	count_withdrawals
0	2018-05-25	1093	NaN	NaN	NaN	3.15	0	0	1
1	2017-03-18	1093	7.50	6.50	100.0	NaN	72	1	0
2	2017-03-17	1093	48.63	38.51	1000.0	NaN	93	2	0
3	2018-12-10	2622	36.29	32.99	NaN	NaN	44	0	0
4	2018-12-09	2622	101.19	102.56	NaN	NaN	76	0	0

Table 2 - transactions table

```
In [6]: campaigns = pd.read_csv('campaigns.csv')
        campaigns.head()
```

Out[6]:

	start_date	end_date	total_spend	country
0	2017-01-03	2017-01-04	9935.67	NaN
1	2017-01-04	2017-01-06	17696.41	id
2	2017-01-08	2017-01-12	22541.26	in
3	2017-02-13	2017-02-18	745.89	ng
4	2017-05-02	2017-05-03	15338.78	id

Table 3 - campaigns table

4.2 Data Wrangling

The main goal in this section is to clean and trim the data to be ready for descriptive and advanced analysis. so we investigate data to see any anomaly in each table. we first change the format of some columns and then identify null values and decide which ones should be omitted and which ones should be replaced with data.[2]

4.2.1 Data formatting

Many columns in each table were originally formatted as datetime but during the importing, data type has been changed. so, columns which are giving information about date and time need to be formatted.

```
basic information

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 columns):
account                    5000 non-null int64
type                       5000 non-null object
residence                  4698 non-null object
date_joined                5000 non-null object
indication_coupon          5000 non-null int64
first_deposit_amount       3959 non-null float64
first_deposit_date         3964 non-null object
first_transaction_amount   2606 non-null float64
first_transaction_date     2608 non-null object
balance_amount             3964 non-null float64
dtypes: float64(3), int64(2), object(5)
memory usage: 390.7+ KB
None
```

```
number of null values in each column

account                        0
type                           0
residence                    302
date_joined                    0
indication_coupon              0
first_deposit_amount        1041
first_deposit_date          1036
first_transaction_amount    2394
first_transaction_date      2392
balance_amount              1036
dtype: int64
```

Figure 2 – basic information about each table and number of null values in each column

```
#clients
#change the column name of residence to country
clients.rename(columns={'residence': 'country'}, inplace = True)

#change the date format from object to datetime
clients['date_joined'] = pd.to_datetime(clients['date_joined'], format="%Y-%m-%d")
clients['first_deposit_date'] = pd.to_datetime(clients['first_deposit_date'], format="%Y-%m-%d")
clients['first_transaction_date'] = pd.to_datetime(clients['first_transaction_date'], format="%Y-%m-%d")
print("clinets basic information after formatting------------------- \n")
print( clients.info())

#transactions
#change the date format from object to datetime
transactions['transaction_date'] = pd.to_datetime(transactions['transaction_date'], format="%Y-%m-%d")
print("transactions basic information after formatting------------------- \n")
print(transactions.info())

#campaigns
#change the date format from object to datetime
campaigns['start_date'] = pd.to_datetime(campaigns['start_date'], format="%Y-%m-%d")
campaigns['end_date'] = pd.to_datetime(campaigns['end_date'], format="%Y-%m-%d")
print("campaigns basic information after formatting------------------- \n")
campaigns.info()
```

Figure 3 – codes run for data formatting

[2] Only some of the tables are shown here. For more info about data wrangling please open the Jupyter Notebook

4.2.2 Identification of missing values

Missed values need to be analyzed with more details. so, each table's missed values has been visualized and then based on basic information, some of them will be omitted.

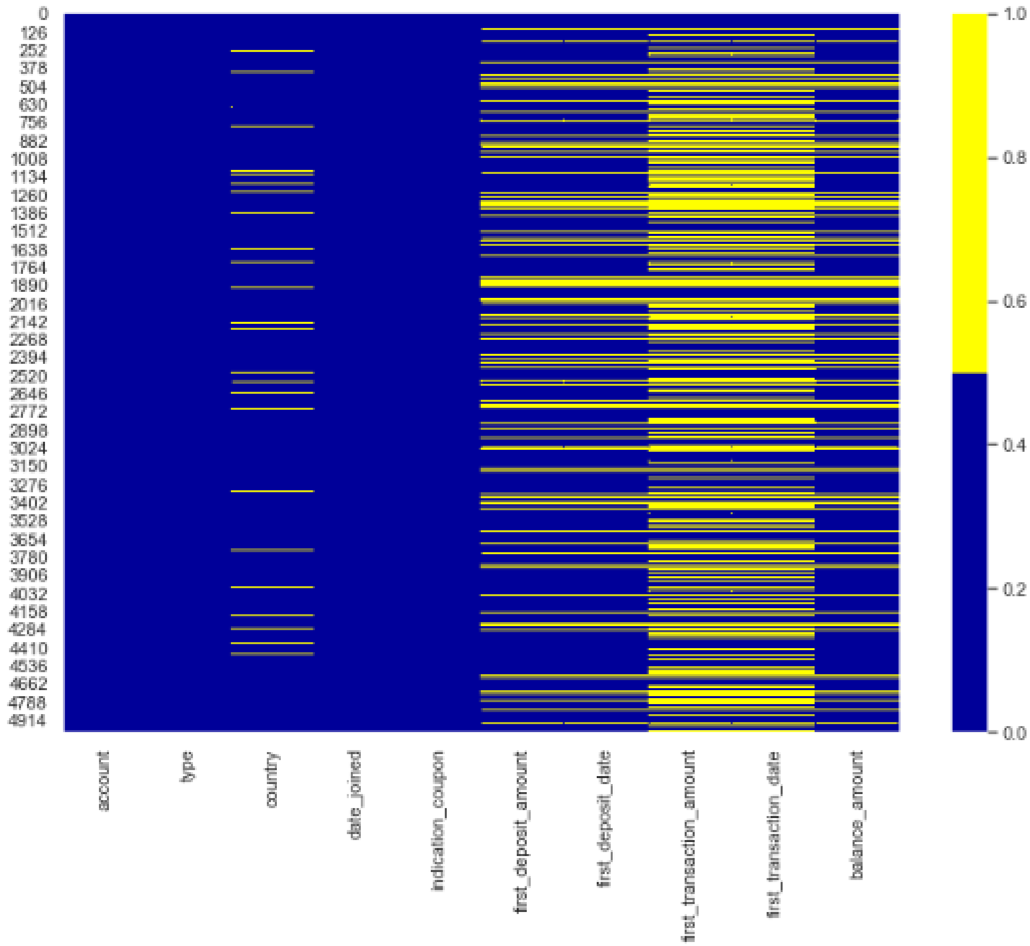

Figure 4 – heatmap of null values in clients table

As we can see, 6 different columns contain Null values. first of all, it seems that some accounts do not belong to any country. but we have seen that in the campaign table, there were some campaigns without any country name and they are categorized as global campaigns. so, it is logical to categorize those accounts without country name as global. then when investigating other null values, we understood that they are necessary to be left omitted because this lack of info will be used in next sections.

4.3 Exploratory Data Analysis[3]

In this section, we firstly categorize our data into our three tables. then, based on characteristics of each table, we conduct some descriptive analysis to broaden our knowledge about our data and reach to some conclusions about the dataset.

4.3.1 clients table

At first, we query the table and generate three tables, namely clients_just_joined, clients_with_deposit and clients_normal.

With the first table(clients_just_joined), we aim to find customers who have just created an account and never decided to buy and sell on the site. they did not even make a deposit to show they are serious. so, they are just passive users on the site.

The second group (clients_with_deposit) seems to be more serious than the first ones as they have paid the deposit. but unfortunately, like the first group, they did not participate actively in buying and selling. they maybe joined the site due to some advertisements during the campaigns and then gave up becoming an active user. so, to better understanding the data, they are categorized as a distinct group.

And finally, customers who buy and sell on the site are separated to investigate more about them as they are the one who make profit.

4.3.1.1 type

These tables have been compared in type, country and indication coupon. Below, you can see more comparison in type.

	type	total_total	total_just_joined	total_with_deposit	total_normal
0	type_6	3238	1	1271	1966
1	type_1	1630	969	63	598
2	type_3	50	31	5	14
3	type_2	35	21	3	11
4	type_4	32	13	6	13

Table 4 – categorized table of three different clients based on type groups

[3] some tables were too long to be shown here. So, they were truncated. To see full table, please open the Jupyter Notebook related to this project

It can be seen that the majority of customers belong to the type_6 and more than half of them are buying and selling on the site and surprisingly, only 1 of the customers in this group has never made a deposit and just joined the site. However, the most fascinating information gained in this table is that around two third of customers who were

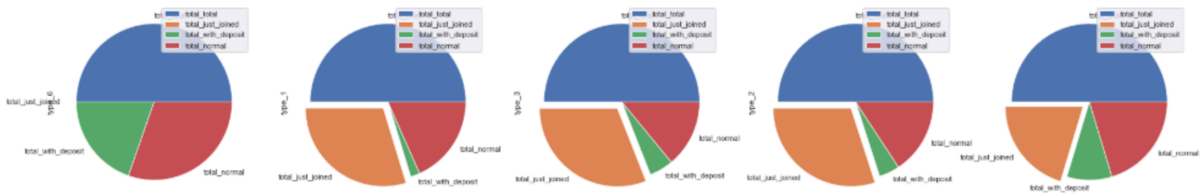

Figure 5 – categorized pie chart of three different clients based on type groups

categorized as type 1 are just joined the site and never start buying or selling. Other customer types seem to have followed the same pattern as type one, as the number of active users among them are less than passive ones.

4.3.1.2 country

	country	total	total_just_joined	total_with_deposit	total_normal
0	id	1667	340	379	948
1	br	438	86	105	247
2	ng	284	80	75	129
3	ru	274	37	53	184
4	vn	170	47	35	88
5	in	141	36	39	66
6	gb	134	18	44	72
7	za	105	29	32	44
8	lk	97	20	10	67
9	ua	77	13	24	40

Table 5 - categorized table of three different clients based on countries

The table above have followed the same pattern as type, showing top 10 popular countries. the country "id" is the home of almost a third of all our customers. more than half of our customers living in "id" actively buy and sell products and make a great portion of site's income. The second most popular country is "br" with 438 customers, which is far less than 1667 of "id"'s customers. then there are "ng" and "ru" in the third and fourth place respectively.

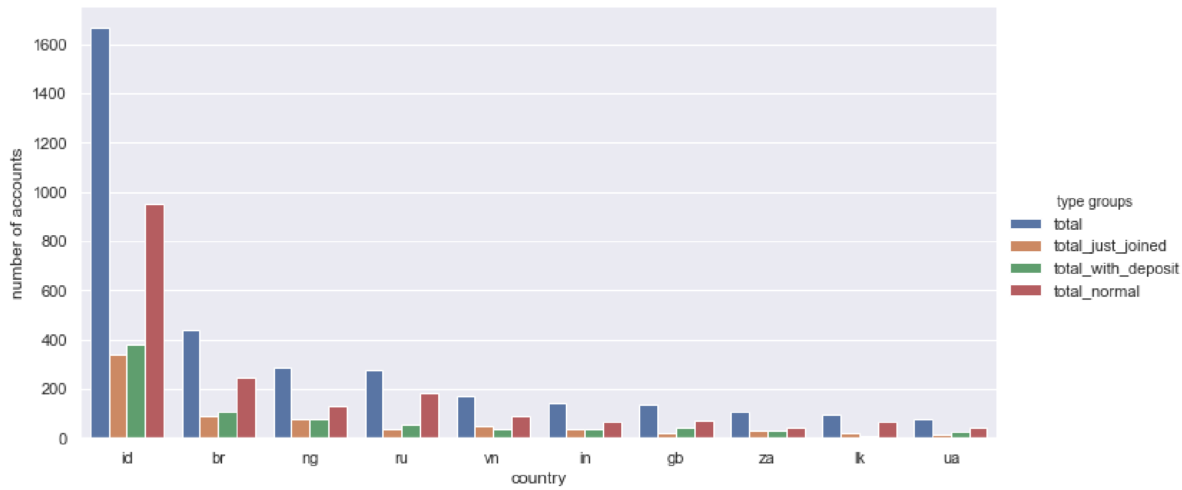

Figure 6 - categorized bar chart of three different clients based on countries

4.3.1.3 Coupon indicator

Among 5000 customers, 4229 of them are using the coupon number 308, which is by far the most popular coupon among customers.

	indication_coupon	total	total_just_joined	total_with_deposit	total_normal
0	308	4229	825	1194	2210
1	41	61	19	12	30
2	116	48	26	11	11
3	36	20	5	2	13
4	181	18	8	1	9
5	48	14	3	3	8

Table 6 - categorized table of three different clients based on indication coupon

4.3.2 transactions table

The main goal of all these tables is to calculate and boost up the profit. and the profit has a direct relationship with total amount of buy and sell (1% of the total). so, in the first step, we sum up total expenditures on the site over the past two years. then these numbers will be categorized based on year and finally season. we can also see the total number of contracts made and total number of deposits as well as withdrawals.

4.3.2.1 total expenditure

	years	sum_total_buy	sum_total_deposits	sum_total_sell	sum_total_withdrawals
0	2017	30,503,986	22,915,212	23,883,429	946,791,604
1	2018	46,685,312	33,632,648	35,630,821	525,012,866
2	Total	77,189,298	56,547,860	59,514,249	nan

Figure - Total amount of money transferred in four different categories over 2 years

In 2018, total amount of buy, sell and deposits has been increased considerably, compared to the 2017 figures. However, the total amount of withdrawals has declined and almost reached to its half just over a year. it shows that in 2018, customers were more likely to keep their money in site rather than making withdrawals, which is a plus sign for site.

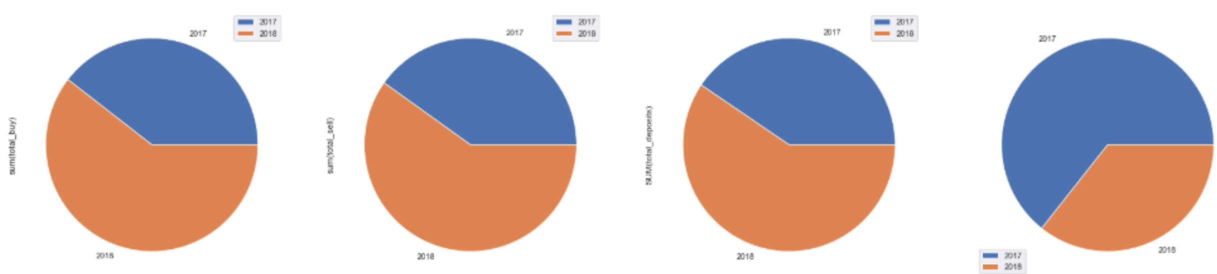

Figure 7- Comparison between expenditures in 2017 and 2018.
From left to right: total buy, total deposit, total sell, total withdrawals

4.3.2.2 total counts

	years	sum_count_contracts	sum_count_deposits	sum_count_withdrawals
0	2017	6,524,113	4,299	956
1	2018	2,882,585	6,455	2,162
2	Total	9,406,698	10,754	3,118

Table 7- Total number of contracts, deposits and withdrawals over 2 years

by looking at table above, it is clear that while total amount of buy and sell have been risen during the year 2018, total number of contracts in this year is less than half of its figure in 2017. this shows that although less contracts have been made, customers have spent more money on each one, which can be because of campaigns.

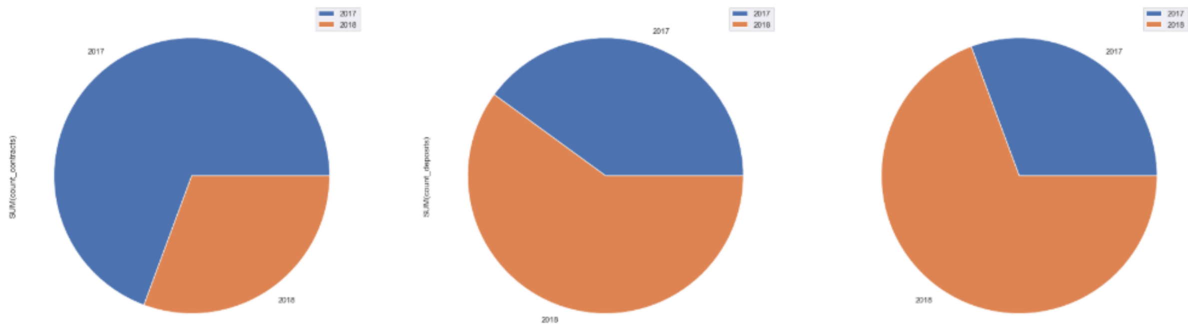

Figure 8 – from left to right: total number of contracts, deposits, withdrawals in 2017 and 2018

4.3.2.3 transactions in each season

By dividing the transactions table into seasons, we have a chance to analyse meticulously. So, by looking at total buy, it is clear that the total buy and sell at autumn 2017 has never been repeated throughout the period shown. As well as this, contracts made during this season and year (around 5.5 millions) are more than total contracts made throughout the two years.

year_season	total_buy	total_sell	total_deposits	total_withdrawals	count_contracts	count_deposits	count_withdrawals
2017- winter	4,116,740	3,155,187	4,590,144	135,733,583	207,739	792	121
2017- spring	4,254,767	3,251,559	4,821,667	228,482,330	336,049	1,042	219
2017- summer	8,764,718	6,763,547	5,515,405	276,038,424	460,212	1,070	248
2017-autumn	13,367,762	10,713,136	7,987,997	306,537,267	5,520,113	1,395	368
2018- winter	13,728,908	10,496,727	8,572,915	190,895,177	663,059	1,391	355
2018- spring	10,098,762	7,646,455	8,656,570	144,775,188	678,101	1,579	395
2018- summer	11,071,387	8,436,031	9,360,418	110,505,890	891,800	1,962	763
2018-autumn	11,786,255	9,051,607	7,042,744	78,836,611	649,625	1,523	649

Table 8– transactions in each season

By looking at the chart of total buy, sell and deposits, it can be seen that at first, the total amount of buy and sell remain relatively still for the first three months of the 2017. However, this figures started to rise sharply over the next 9 months, reaching more than 13 million dollars buy and around 10 million dollars sell by the end of the year 2017 and the first season of the year 2018. While there was a major fall in spring in 2018, the total buy and sell again continued to grow to reach around 11 million and 8 million respectively.

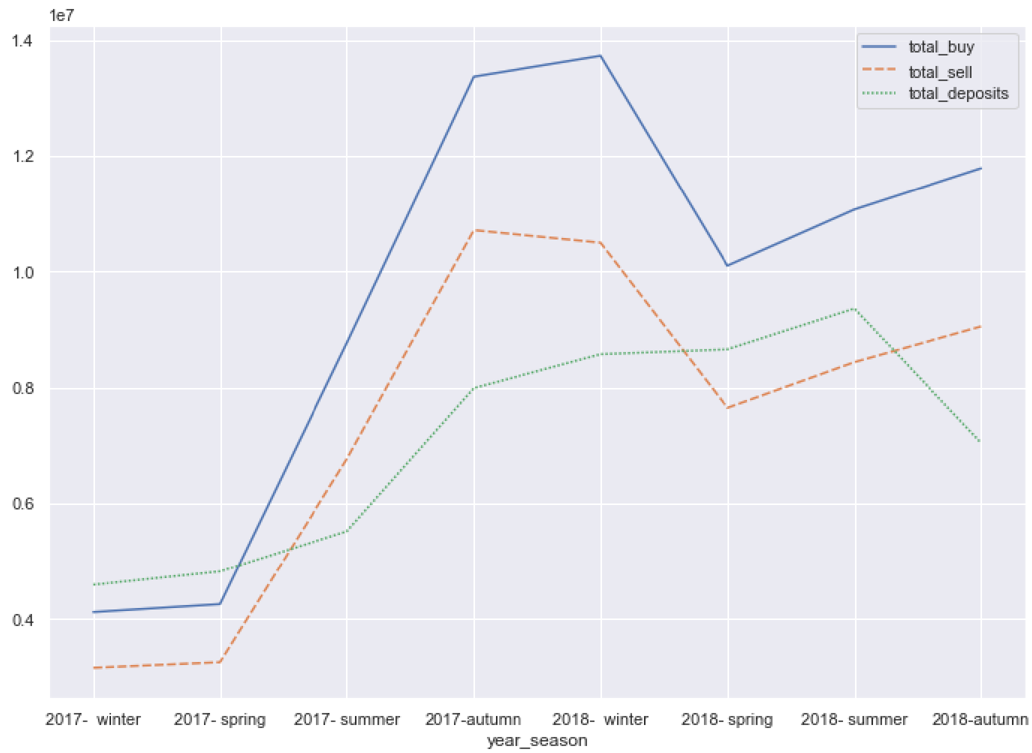

Figure 9– Transactions in each season – total buy, total sell and total deposits

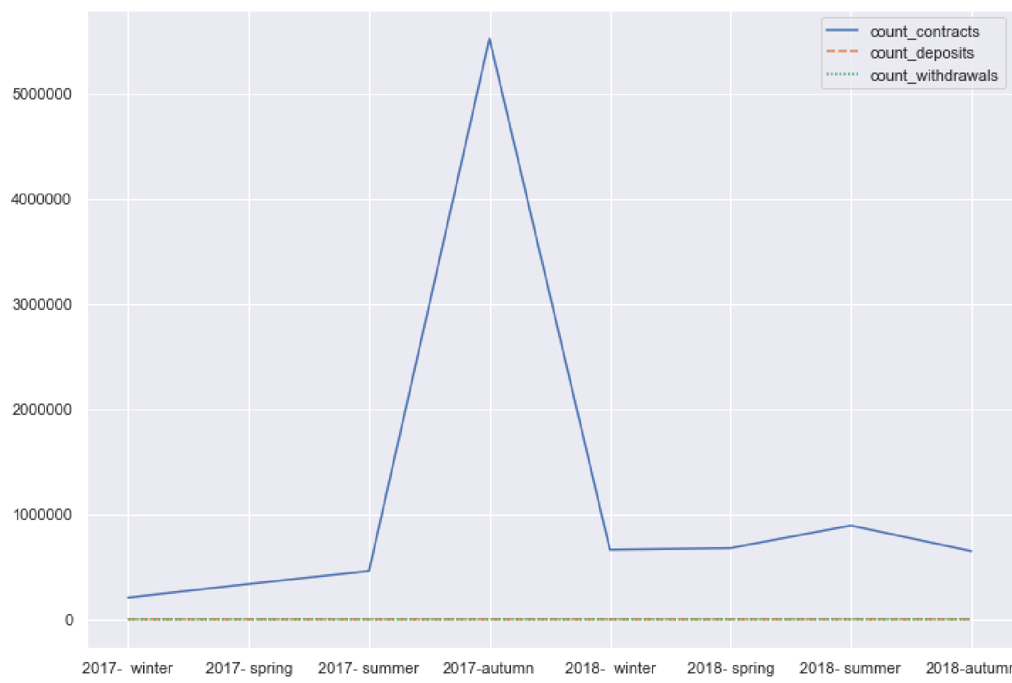

Figure 10 – total number of contracts deposits and withdrawals in each season

4.3.3 campaigns

```
-------------------
 number of campaigns in each country:
          country
id          14
ng           4
br           4
ru           3
ua           2
global       2
gh           2
gb           1
za           1
ca           1
py           1
in           1
co           1
ph           1
de           1
ie           1
ma           1
mg           1
it           1
-------------------
```

	years	total_spent
0	2017	320,444
1	2018	195,765
2	Total	516,209

year_season	total_spend
2017- winter	50,919
2017- spring	57,738
2017- summer	138,164
2017-autumn	73,623
2018- winter	85,871
2018- spring	71,953
2018- summer	16,662
2018-autumn	21,279

Table 9 – information about campaigns

Campaigns table is somehow similar to transactions. So, the same method and tools will be done to extract information.

at first, we can see that 14 out of 43 (around a third) campaigns was held in a country called "id", which shows how important this country is for our site. Around 320,000 dollars was invested on campaigns during 2017, but the manager team decided to decrease the expenditures to less than 200,000 in the next year. But the most important table here is the amount of money spent as campaign expenditure in each season. During the summer 2017, the company decided to run most of its campaign worth around 140,000 US dollars, which was the highest over the period shown. After that, total campaign expenditures suddenly fell and had reached to its lowest point by the next summer in 2018.

The next figure is a comparison between total buy and sell divided by 100 (because 1% of each buy or sell is net profit of the site) and campaign expenditure in each season. It is clear that during the first month, company have invested money on campaigns to raise the revenue, which seems be successful, and then when total buy and sell raised enough, campaign budget decreased. At first, it affected the total revenue, by falling from around 140,000 from total buy and around 100,000 from sell to 100,000 and 80,000 respectively, but it remained almost unchanged and even saw improvements for the rest of the period.

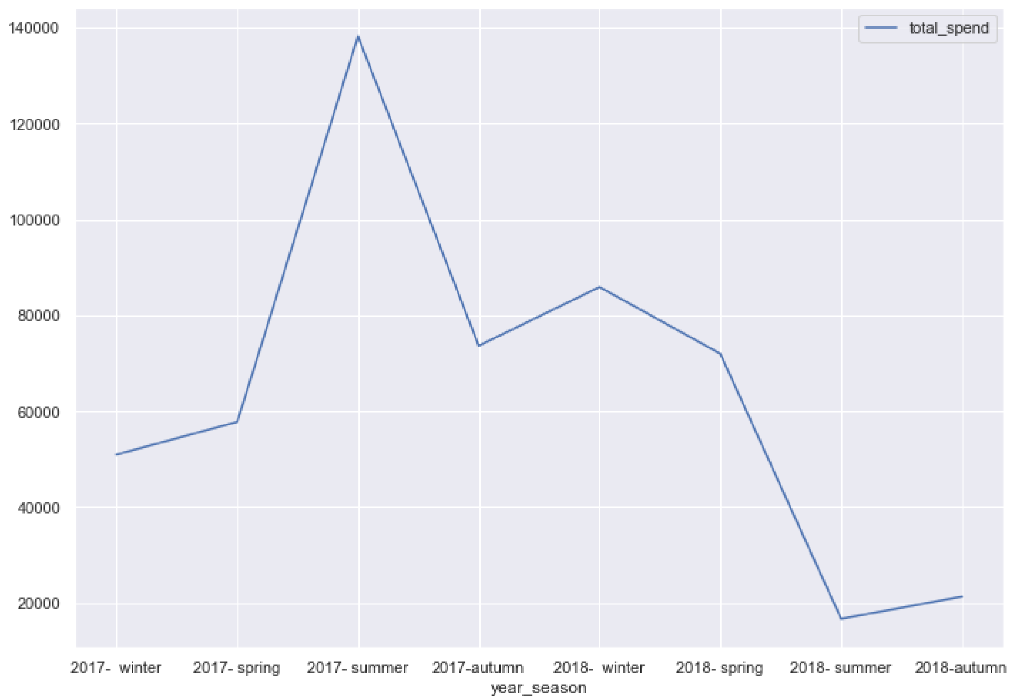

Figure 11 – total spend on each campaign in each season

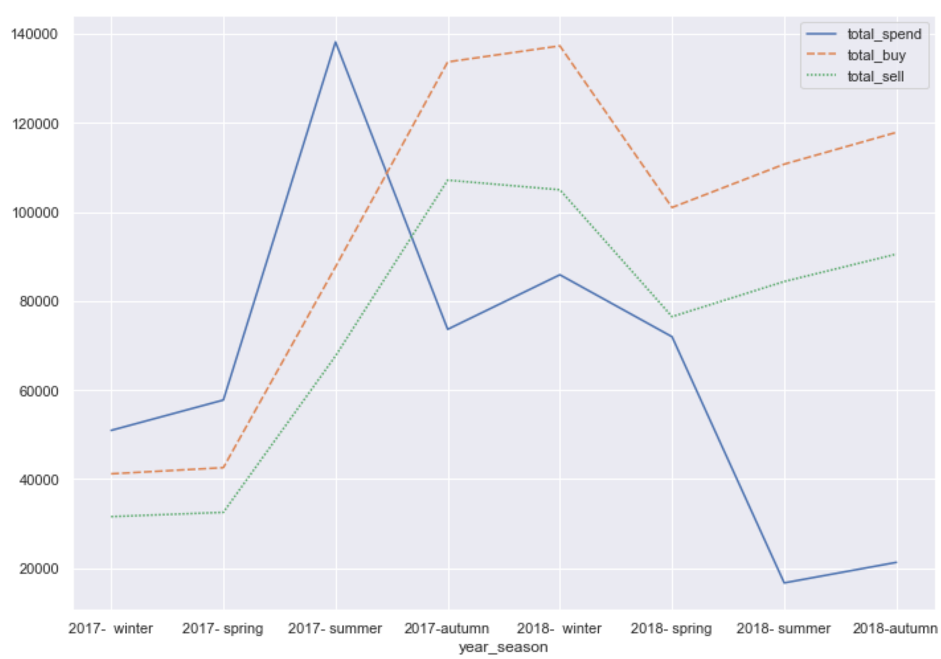

Figure 12 – comparison between total spend and net revenue from total buy and total sell

4.4 More Advanced Analysis[4]

After analyzing each table separately, now we want to merge tables to see how we can analyze better. The main goal is to see whether campaigns were successful enough or not. So, the first step is to analyze this hypothesis by categorizing each transaction by campaigns that had direct or indirect effect on transactions. Then we need to analyze the impact of campaign not only on each transaction, but also on each client. So, categorized customers need to be extracted to see who is actually visiting the site, who make profit for us and who are potential customers for earning money in the future. To avoid any misunderstanding, each step in hypothesizes mentioned above is explained thoroughly.

4.4.1 categorizing transactions related to each campaign based on transaction_date

at first, each campaign is enumerated. then a merged table of clients and transactions is queried to gain more insight on each transaction by knowing what is the account number, country, type of customer and coupon indicator of each transaction. after that, we try to investigate which transactions happened during the campaign time. When we know exactly which campaign have brought which customers, then the performance of each campaign can be calculated. then, there is a hypothesis saying that if an account joined the site or have at least a transaction during a campaign, the next transactions can be considered to be related to that campaign. so we categorize them as campaign_number_after in a separate column to investigate the performance of each campaign. in the end we can exactly see which campaigns had more great impact on our transactions during the past two years.

4.4.1.1 allocating a number to each campaign

	start_date	end_date	total_spend	country	year_season	campaign_number
0	2017-01-03	2017-01-04	9935.67	global	2017-winter	1
1	2017-01-04	2017-01-06	17696.41	id	2017-winter	2
2	2017-01-08	2017-01-12	22541.26	in	2017-winter	3
3	2017-02-13	2017-02-18	745.89	ng	2017-winter	4
4	2017-05-02	2017-05-03	15338.78	id	2017-spring	5

Table 10 - allocating a number to each campaign

[4] some tables were too long to be shown here. So, they were truncated. To see full table, please open the Jupyter Notebook related to this project

4.4.1.2 creating a new table consist of clients and transactions

	transaction_date	account	type	country	indication_coupon	total_buy	total_sell	total_deposits	total_withdrawals	count_contracts
0	2018-05-25	1093	type_1	ru	13	NaN	NaN	NaN	3.15	0
1	2017-03-18	1093	type_1	ru	13	7.50	6.50	100.0	NaN	72
2	2017-03-17	1093	type_1	ru	13	48.63	38.51	1000.0	NaN	93
3	2018-12-10	2622	type_1	ng	308	36.29	32.99	NaN	NaN	44
4	2018-12-09	2622	type_1	ng	308	101.19	102.56	NaN	NaN	76

Table 11- creating a new table consist of clients and transactions

4.4.1.3 which campaign, which transaction

try to see which transactions and how many of them have happened during the campaign time.

campaign_number	number_of_transactions
0	35284
26	136
33	119
16	110
27	94

Table 12 – transactions which have happened during each campaign

4.4.1.4 grouped campaigns

To gain a broader view, we grouped campaigns which had a campaign number based on total buy and sell. then a new table was created consist of campaign information and then sort all of them based on total buy. As we can see below, campaign number 33 located in "id" had the greatest effect on total amount of buy and sell in our table. other campaigns are sorted as well.

	campaign_number	total_buy	total_sell	total_spend	country	year_season	number_of_transactions_during_each_campaign
0	33	528492.57	422144.20	9192.02	id	2018- winter	119
1	16	523279.94	408732.33	13731.69	id	2017- summer	110
2	26	384651.70	294318.64	15020.37	id	2017-autumn	136
3	27	311620.18	252865.05	5533.58	id	2018- winter	94
4	24	307850.18	240802.23	5491.55	id	2017-autumn	75
5	28	255082.71	199387.61	11466.04	id	2018- winter	54
6	32	231675.53	178235.27	12906.71	id	2018- winter	70
7	13	229420.45	174151.22	26143.24	id	2017- summer	87
8	41	196755.07	145537.21	1745.10	id	2018-autumn	70
9	34	161245.19	121779.60	19928.51	id	2018- spring	88

Table 13 – grouped campaigns with number of transactions during each campaign

4.4.1.5 customer transactions after the campaigns

The hypothesis here is that if a customer join the club, the next transactions are indirectly related to that campaign. so, we firstly found out which accounts were active during the campaigns. then dedicate campaign number to transactions which have happened after the related campaign in a column named campaign_number_after. in the end, the number of transactions happened after each campaign can be counted.

	number_of_transactions
campaign_number_after	
0	26523
34	1411
33	923
16	914
26	848

Table 14 – transactions which have happened after each campaign

4.4.1.6 grouped campaigns after

	total_buy	total_sell	total_spend	country	year_season	campaign_number	num_of_tr_after_each_camp
0	4593704.61	3505996.17	13731.69	id	2017-summer	16	914
1	4065571.53	3175790.87	1745.10	id	2018-autumn	41	833
2	3599739.68	2803737.65	15020.37	id	2017-autumn	26	848
3	3332225.57	2481085.35	9192.02	id	2018-winter	33	923
4	2911009.74	2229592.37	11466.04	id	2018-winter	28	630
5	2907832.81	2254291.41	26143.24	id	2017-summer	13	560
6	1736151.30	1308946.45	19928.51	id	2018-spring	34	1411
7	1431323.14	1095744.93	12906.71	id	2018-winter	32	326
8	1116435.49	864791.30	15338.78	id	2017-spring	5	350
9	762247.54	554428.61	11581.60	ng	2018-spring	36	143

Table 15– grouped campaigns with number of transactions after each campaign

4.4.2 categorizing each client based on campaigns

At first, we intend to see which accounts have joined, made a deposit or made their first transaction during the campaigns to analyze the impact of each campaign. we also looked for accounts which have at least bought or sold a single product on the site during the campaigns. Then we counted the number of transactions for each account number to categorise accounts based on their importance on total number of transactions they have got. The same algorithm was run to calculate total amount of money bought or sold buy each account number. now we can analyse whether campaigns were successful to attract profitable customers or not.

4.4.2.1 number of accounts which have joined during each campaign

	campaign_number	number_of accounts_date_joined
0	0	4865
1	26	20
2	33	15
3	41	12
4	27	11
5	13	11

Table 16 - number of accounts which have joined during each campaign

4.4.2.2 number of accounts which have made the first deposit during each campaign

	campaign_number	number_of accounts_first_deposit
0	0	4885
1	33	13
2	26	12
3	13	11
4	41	10
5	31	9

Table 17 - number of accounts which have made the first deposit during each campaign

4.4.2.3 number of accounts which have made the first transaction during each campaign

	campaign_number	number_of accounts_first_transaction
0	0	4929
1	33	10
2	13	10
3	34	7
4	26	6
5	16	6

Table 18- number of accounts which have made the first transaction during each campaign

4.4.2.4 accounts which have at least made a transaction during each campaign

the ones which have done a purchase are 1 and the others are 0

	experienced_or_not	number_of_accounts_experienced_campaign
0	0	4645
1	1	355

Table 19- accounts which have at least made a transaction during each campaign

4.4.2.5 number of orders each account made

	account	type	country	number_of_transactions
1432	1103	type_1	lk	303
272	1305	type_1	id	295
2791	1008	type_1	id	291
2650	2997	type_6	ru	291
2411	1113	type_1	ru	279

Table 20- number of orders each account made

4.4.2.6 total amount of money bought and sold by each account

	account	total_buy	total_sell	type	country
0	3479	4220269.60	3251589.47	type_6	id
1	3265	4065269.07	3049223.58	type_6	id
2	5412	4064804.85	3073714.90	type_6	id
3	3710	2598161.48	1928259.02	type_6	id
4	4145	1780723.77	1379848.07	type_6	gh

Table 21- total amount of money bought and sold by each account

4.4.2.7 merged table

account	type	country	date_joined_campaign_number	first_deposit_campaign_number	first_transaction_campaign_number	campaign_experience	number_of_transactions	total_buy	total_sell
4003	type_6	id	26	26	26	1	161	1153347.46	897037.32
2768	type_6	id	2	2	0	1	7	157149.77	120791.94
4478	type_6	id	33	33	33	1	60	136058.29	102934.72
4436	type_6	br	31	31	31	1	20	121605.88	94444.44
4662	type_6	id	34	34	34	1	1	41694.23	31676.62
4495	type_6	id	33	33	33	1	2	34819.00	26333.09
4479	type_6	id	33	33	33	1	2	32538.22	24905.61
3981	type_6	id	26	26	26	1	2	30728.88	20753.36
4648	type_6	id	34	34	34	1	26	26000.70	20135.31
5804	type_6	br	42	42	0	1	12	25570.69	18209.83

Table 22– merged table of categorized clients based on campaigns

5 Discussions & Conclusions

In this section, our main goal is to answer questions discussed before and try to find out which hypothesis were right or right. So, we highlight each question and will answer them with details. Then we discuss insights discovered during the analysis which might be helpful for further strategic decision makings.

5.1 questions to answer

5.1.1 How was the performance of campaigns during the past two years?

This question was answered numerous time during the result section, but to be more specific, we can see that campaigns had positive effects on boosting the performance at some points. Firstly, by looking at figure 13, it can be deducted that the investments made during the first summer of the site paid back and total buy and sell had increased by the end of that year (2017). when campaign expenditures were declined in the first half of the year 2018, total revenue was decreased at first too. All these dependencies showed a positive relationship between total spend on campaigns and total buy and sell on the site.

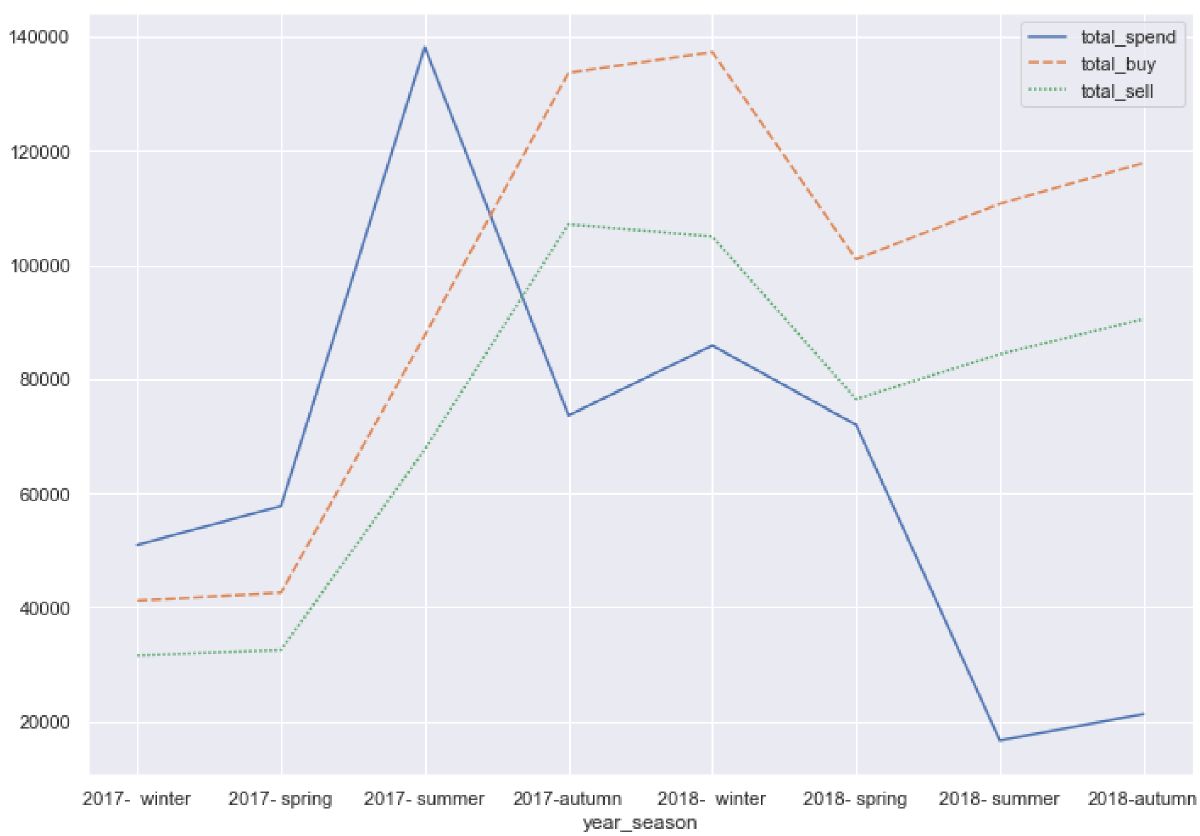

Figure 13 - comparison between total spend and net revenue from total buy and total sell

5.1.2 When a campaign was run, how did the customers behave?

To gain more insight, we categorize each campaign based on transactions happened during and after campaigns. Surprisingly, just a small number of accounts were directly or indirectly (based on our hypothesis) affected by campaigns, which was disappointing at first but still shows that each campaign has brought how many customers, from which country, and how much money was earned by each campaign. In this way, a full analysis can be done to examine the performance of each campaign to see which ones were more successful.

campaign_number	total_buy	total_sell	total_spend	country	year_season	number_of_transactions_during_each_campaign
33	528492.57	422144.20	9192.02	id	2018- winter	119
16	523279.94	408732.33	13731.69	id	2017- summer	110
26	384651.70	294318.64	15020.37	id	2017-autumn	136
27	311620.18	252865.05	5533.58	id	2018- winter	94
24	307850.18	240802.23	5491.55	id	2017-autumn	75
28	255082.71	199387.61	11466.04	id	2018- winter	54
32	231675.53	178235.27	12906.71	id	2018- winter	70
13	229420.45	174151.22	26143.24	id	2017- summer	87
41	196755.07	145537.21	1745.10	id	2018-autumn	70
34	161245.19	121779.60	19928.51	id	2018- spring	88

Table 23 – grouped campaigns with number of transactions during each campaign

total_buy	total_sell	total_spend	country	year_season	campaign_number	num_of_tr_after_each_camp
4593704.61	3505996.17	13731.69	id	2017- summer	16	914
4065571.53	3175790.87	1745.10	id	2018-autumn	41	833
3599739.68	2803737.65	15020.37	id	2017-autumn	26	848
3332225.57	2481085.35	9192.02	id	2018- winter	33	923
2911009.74	2229592.37	11466.04	id	2018- winter	28	630
2907832.81	2254291.41	26143.24	id	2017- summer	13	560
1736151.30	1308946.45	19928.51	id	2018- spring	34	1411
1431323.14	1095744.93	12906.71	id	2018- winter	32	326
1116435.49	864791.30	15338.78	id	2017- spring	5	350
762247.54	554428.61	11581.60	ng	2018- spring	36	143

Table 24 – grouped campaigns with number of transactions after each campaign

When we look at these numbers, it is crystal clear that campaigns held in "id" were the most successful campaigns. Although a third of all campaigns was held in this country, but it seems that the management team were right to target this country and can put aside more money in this region of the world.

5.1.3 Is there any relationship between campaigns' dates and clients?

Some accounts were created during campaigns, some of them made a deposit during campaigns and finally some of them have made their first transaction during these campaigns. We even examined which accounts have at least bought or sell products during campaign times. Beside, we have counted number of transactions made by each account and sum up total buy and sell by each account. A table has made consist of these information which can be queried based on needs.

5.2 out of scope insights

During the analysis, some information was gained which was not in any initial question. Here, we will investigate them. In table below, number of transactions happened during and after campaigns held in each country is shown. We can see that id country is by far the most successful country for campaigns.

country	number_of_tr_during_each_camp	number_of_tr_after_each_camp
id	950	7553
br	86	751
ru	53	1194
ng	37	228
gh	4	34
in	4	8
it	3	1
ca	2	1
de	2	130
gb	1	3

Table 25 – total number of transactions during and after campaigns in each country

Most of the transactions are in this country, which can be seen in table below. So, the question here is that whether the number of campaigns held in this country was enough or not?

country	total_buy_during_campaigns	total_sell_during_campaigns	total_buy_after_campaigns	total_sell_after_campaigns
id	3,247,325	2,523,096	27,109,548	20,764,635
ng	84,397	62,383	1,051,537	777,284
br	65,649	50,183	514,268	387,023
ru	38,040	29,115	116,873	83,798
it	14,283	11,416	0	0
ca	4,942	3,074	76	61
gh	1,081	1,043	28,352	23,694
in	165	136	873	671
gb	40	16	562	497
de	27	21	3,146	2,334

Table 26 – total amount of buy and sell during and after campagins in each country

To answer this question, we need to find out which transactions did not happen during campaigns to see potential among customers. In table below, we see that while the

country	total_buy	total_sell	total_deposits
id	21,232,474	16,319,758	17,971,689
ng	2,588,713	1,968,110	2,188,746
br	2,560,632	1,948,239	3,892,405
gb	2,440,336	2,042,439	1,426,883
vn	2,234,566	1,666,407	1,888,413

Table 27 – total amount of buy and sell which have spent on site without any influence of campaigns

company have conducted 14 campaigns on this country, there are still a considerable amount of money that have been transferred on this site without any campaign involvement. To make it more clear, let's compare total amount of money spent as total buy and total sell on this site in each country and compare it with numbers extracted above on a table. We have 4 columns for each country, showing how much money has been spent as total buy and total sell, and how much money was spent without any impact campaigns (means that they are organic customers came to visit site without any direct or indirect effect of campaigns). As we can see, the country "id" is by far the most popular region among our customers. It was something predictable. However, while 14 separate campaign was held in this region, there are still many customers who have visited the site without any campaigns. The difference between total_buy and total_buy_no_campaign is the amount of money earned by campaigns. The more this distance is, the more successful campaigns are. Let's investigate other top 10 countries as well

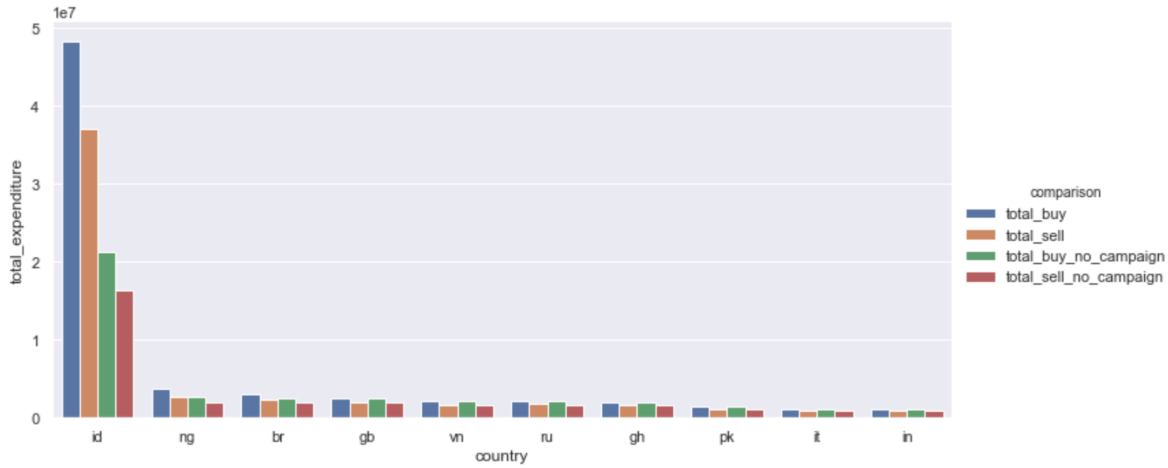

Figure 14 – comparison between total amount of buy and sell and total amount of buy and sell which have joined the site organically – focus on id

it can be seen that campaigns in countries like "ng" and "br" had some minimal effects but almost no effects on other countries can be found, which shows that the strategists and managers need to pay more attention to campaigns held in these regions.

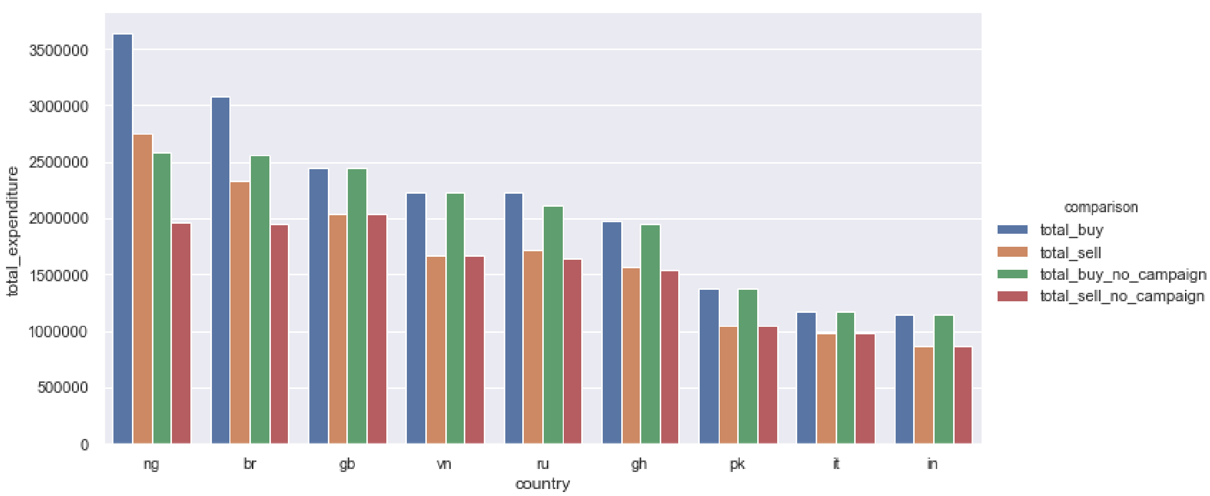

Figure 15 – comparison between total amount of buy and sell and total amount of buy and sell which have joined the site organically – focus on id

to investigate the profitability of each campaign, we have divided the amount of money earned directly and indirectly by campaign by 100 to compare net profit of total buy and total sell with campaign expenditure. We can see that the only profitable campaigns during the past two years were the ones that was run in “id”. Other ones were not able to become even with the cost of campaigns.

country	total_spend	total_buy_during_and_after	total_sell_during_and_after
id	168,463	303,569	232,877
ng	35,351	11,359	8,397
br	47,229	5,799	4,372
ru	40,110	1,549	1,129
it	12,464	143	114
ca	10,095	50	31
gh	7,896	294	247
in	22,541	10	8
gb	16,954	6	5
de	1,683	32	24

Table 28 – comparison between total spend on campaigns in each region and total amount of money earned by these campaigns

In diagrams above, there were some countries where campaigns were unsuccessful as they could not increase transactions enough to be profitable. However, there are some campaigns which they were a total fiasco, as they could not bring any customer at all. They are queried in table below.

	start_date	end_date	total_spend	country	year_season	campaign_number
0	2017-01-03	2017-01-04	9935.67	global	2017- winter	1
5	2017-05-20	2017-05-23	7976.25	gh	2017- spring	6
9	2017-06-28	2017-06-28	9909.73	ua	2017- spring	10
11	2017-07-26	2017-07-27	15395.67	co	2017- summer	12
18	2017-09-21	2017-09-21	2677.16	za	2017- summer	19
19	2017-09-23	2017-09-25	15065.68	global	2017- summer	20
22	2017-10-13	2017-10-15	12718.82	ie	2017-autumn	23
24	2017-11-24	2017-11-26	16456.49	id	2017-autumn	25
29	2018-02-07	2018-02-10	18207.38	ph	2018- winter	30
36	2018-06-24	2018-06-26	3050.58	ma	2018- spring	37
37	2018-06-26	2018-06-28	7974.37	mg	2018- spring	38
39	2018-07-29	2018-08-02	16661.93	py	2018- summer	40
42	2018-10-31	2018-11-02	17392.31	ua	2018-autumn	43

6 Limitations and Suggestions

The analysis in this project has taken place in 4 separate sectors. Firstly, the data was imported as raw data from a database file and then transformed into Pandas dataframes to analyze more. However, SQL queries were used to analyze more. Then the imported data was ready to be cleaned. After that each table was examined separately to extract useful information and gain more insight on data. And finally, tables has been merged and new table were generated to conduct more advanced analysis. During the last phase, numerous tables with the goal of campaign performance was generated which can be used to broaden management knowledge about all the features on the site. Just a small portion of data was explained in Discussion section but there are definitely a wide range of relationships in each table which can be explained and be used for planning the future of the site, but it would be time-consuming and out of scope of this projects.

As well as this, the data and table were purely analyzed by python programming language and SQL, and the only technique was the creativity of the author. However, a wide variety of machine learning algorithm could be used as well. For example, K-means clustering could be used to segment different customers to plan for each cluster of customers separately and define new approaches to make more profit. But the scope of this project would not let such analysis to be done. More interested learners can run such algorithms.

Another point about dataset is that there were lack of information about the usage of deposits and withdrawals one the site. This limitation would not let us analyze these features and related ones (such as balance amount) more deeply.

7 References

[1] firstsource.io

[2] en.wikipedia.org/wiki/Online_shopping

[3] pandas.pydata.org/pandas-docs/