Data Challenge

Sumeet Kotaria

(Background: Using bank statements of various small merchants to cluster the transaction type with an aim to assess the risk profile for loan default)

Index:

1. Data Overview
2. Data Structure
3. Time Series Analysis
4. Time Series Analysis (agg. over banks)
5. Dimension of cash flows
6. Other hypothesized dimensions
7. Clustering

1.

Data Overview

29029 x 11

No. of transactions(rows) x Features (columns)

21 merchants

Unique users

53 bank Accounts

Accounts over 10 banks

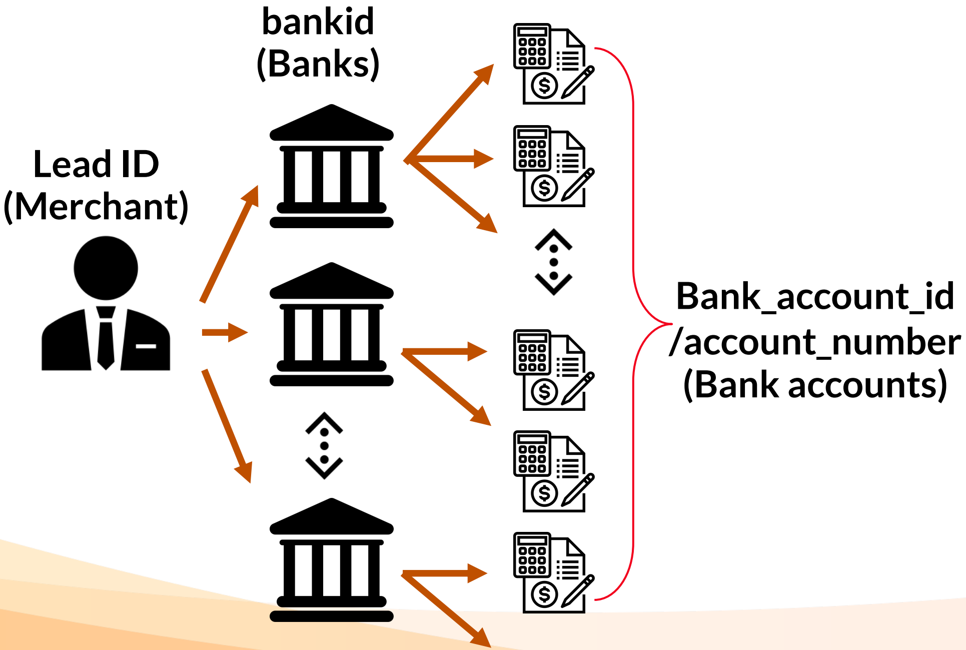
Merchant 321380 has the highest no. of bank accounts

Merchant	No. of bank accounts
321380	8
318465	7
316728	6
321356	5
312745	5
326062	3
314559	2
321146	2
310443	2
323253	2
325330	1
329803	1
328212	1
326050	1
308148	1
325142	1
321671	1
321218	1
314036	1
313082	1
330698	1

2.

Data Structure

Relation b/w Lead ID, bankid, bank_account_id, account_number

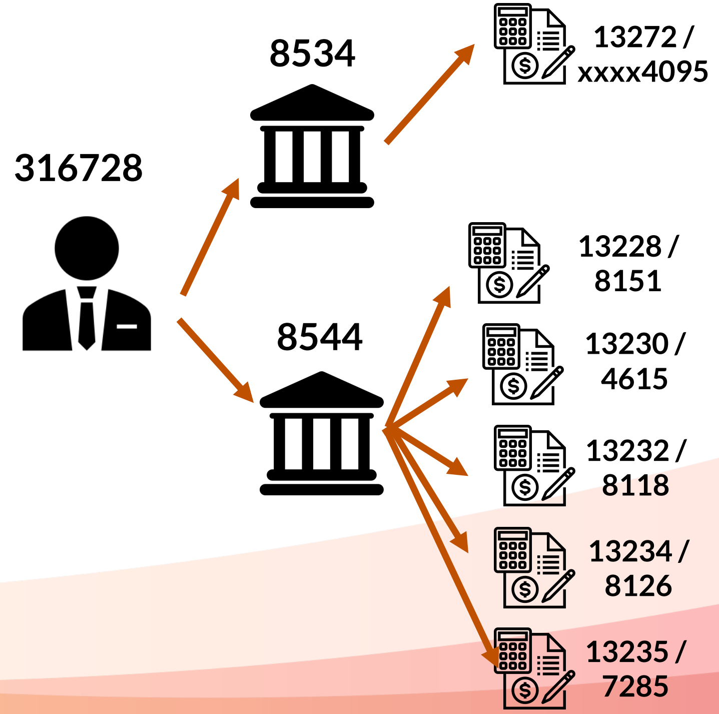


Lead ID
(Merchant)

bankid
(Banks)

Bank_account_id
/account_number
(Bank accounts)

Example



316728

8534

8544

13272 /
xxxx4095

13228 /
8151

13230 /
4615

13232 /
8118

13234 /
8126

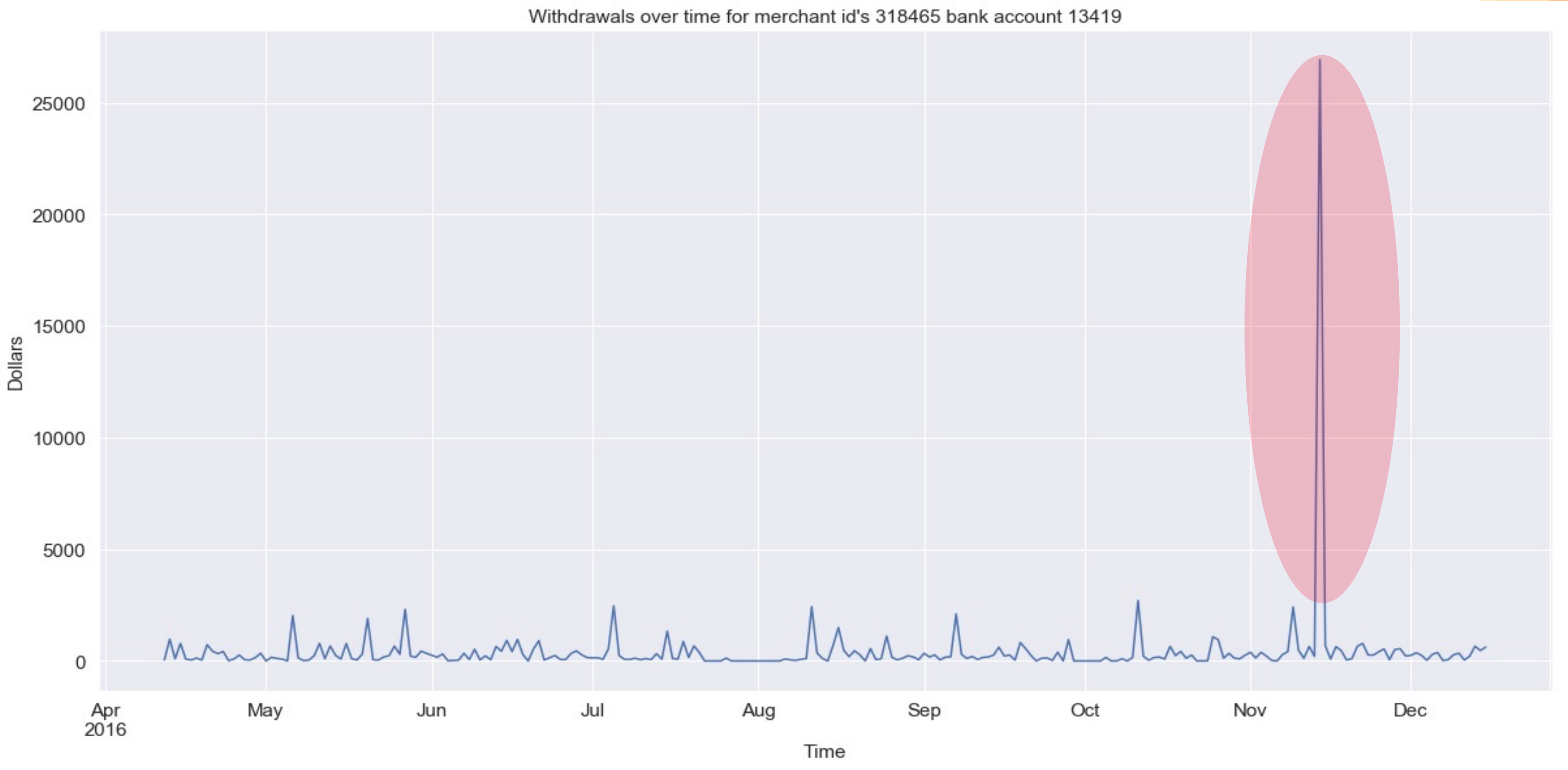
13235 /
7285

3.

Time series analysis

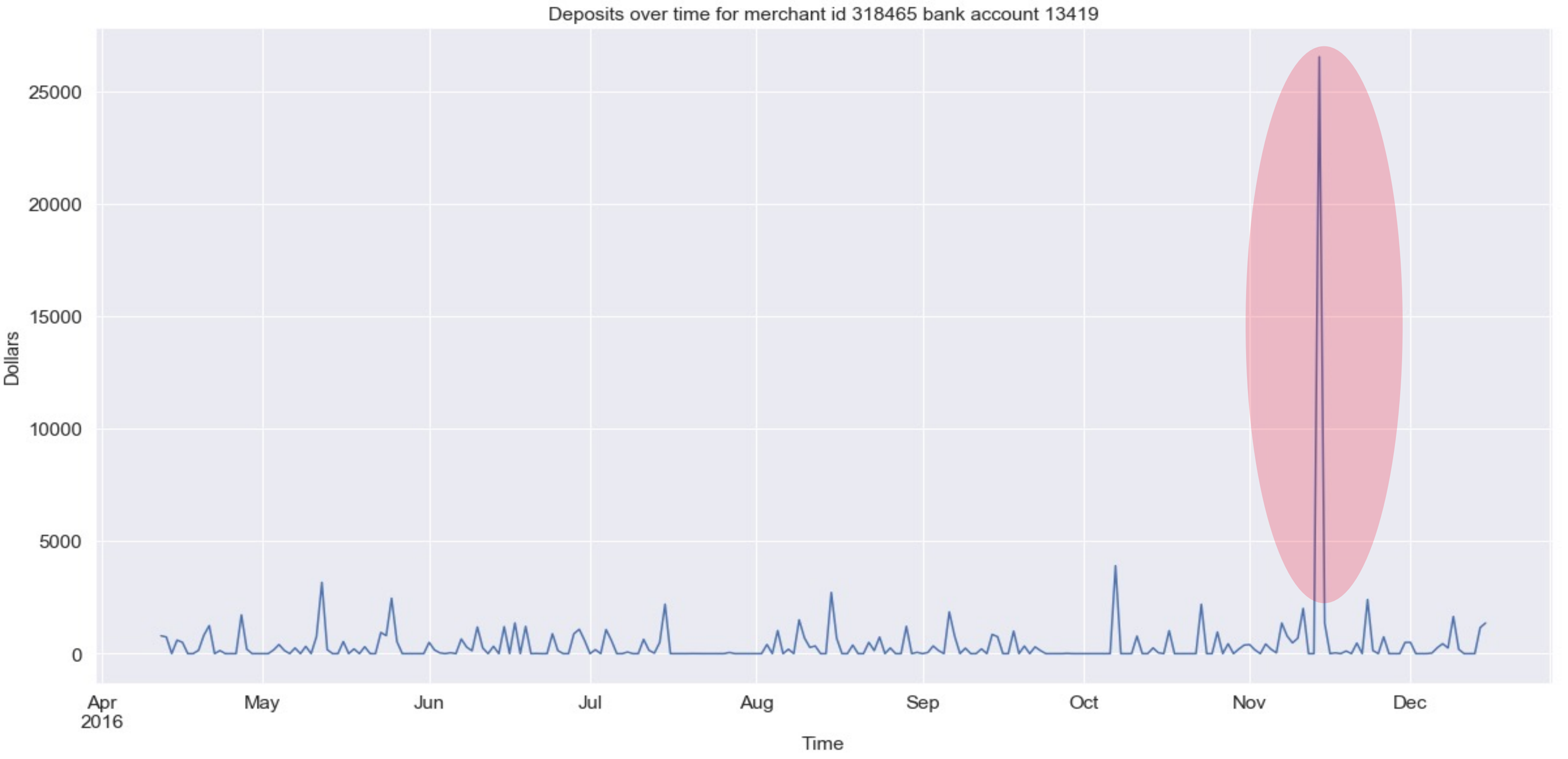
Given Lead ID and Bank Account ID

Debit transaction trend of 318465 merchant’s account id 13419



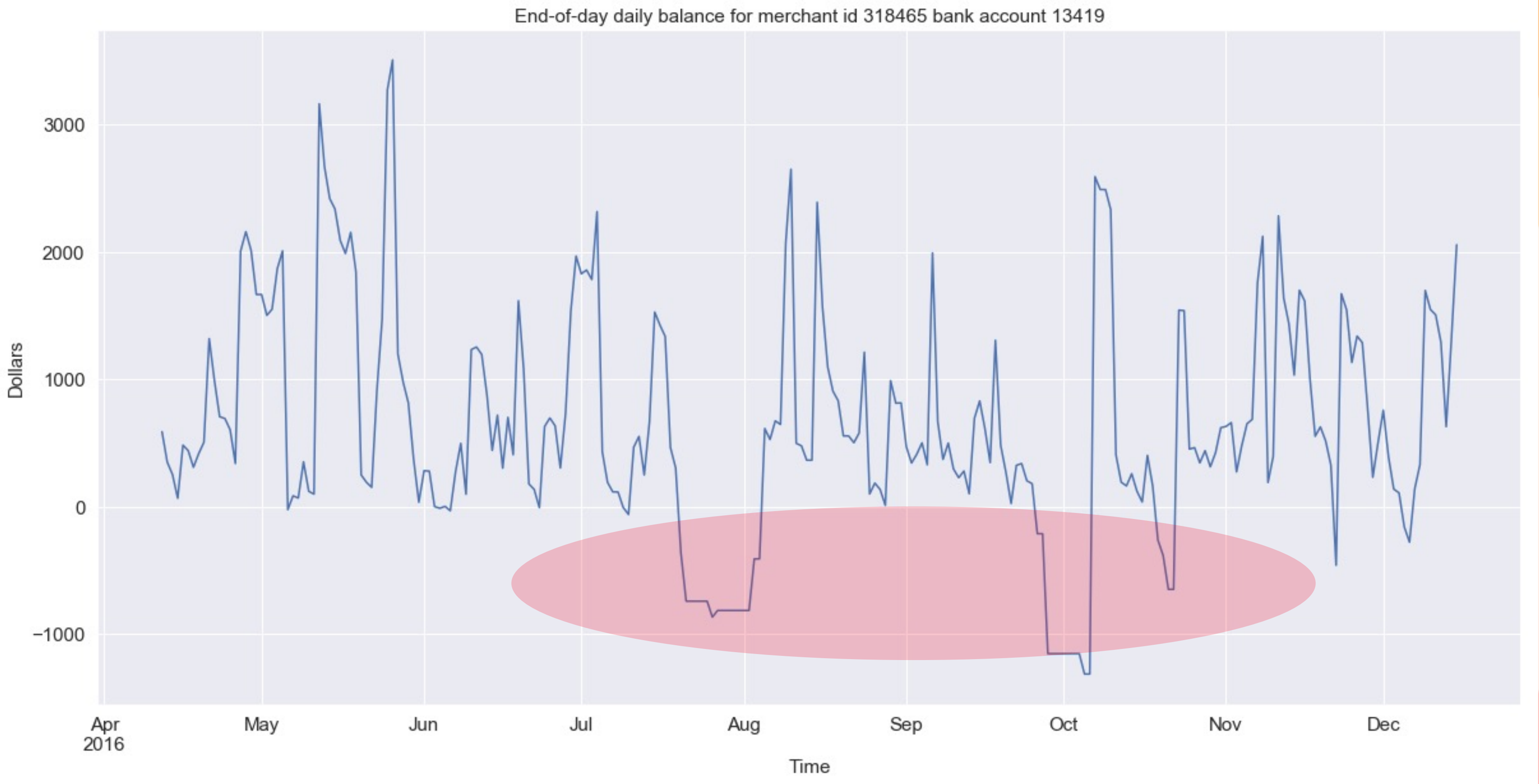
A large withdrawal happened from the account in November 2016

Credit transaction trend of 318465 merchant's account id 13419



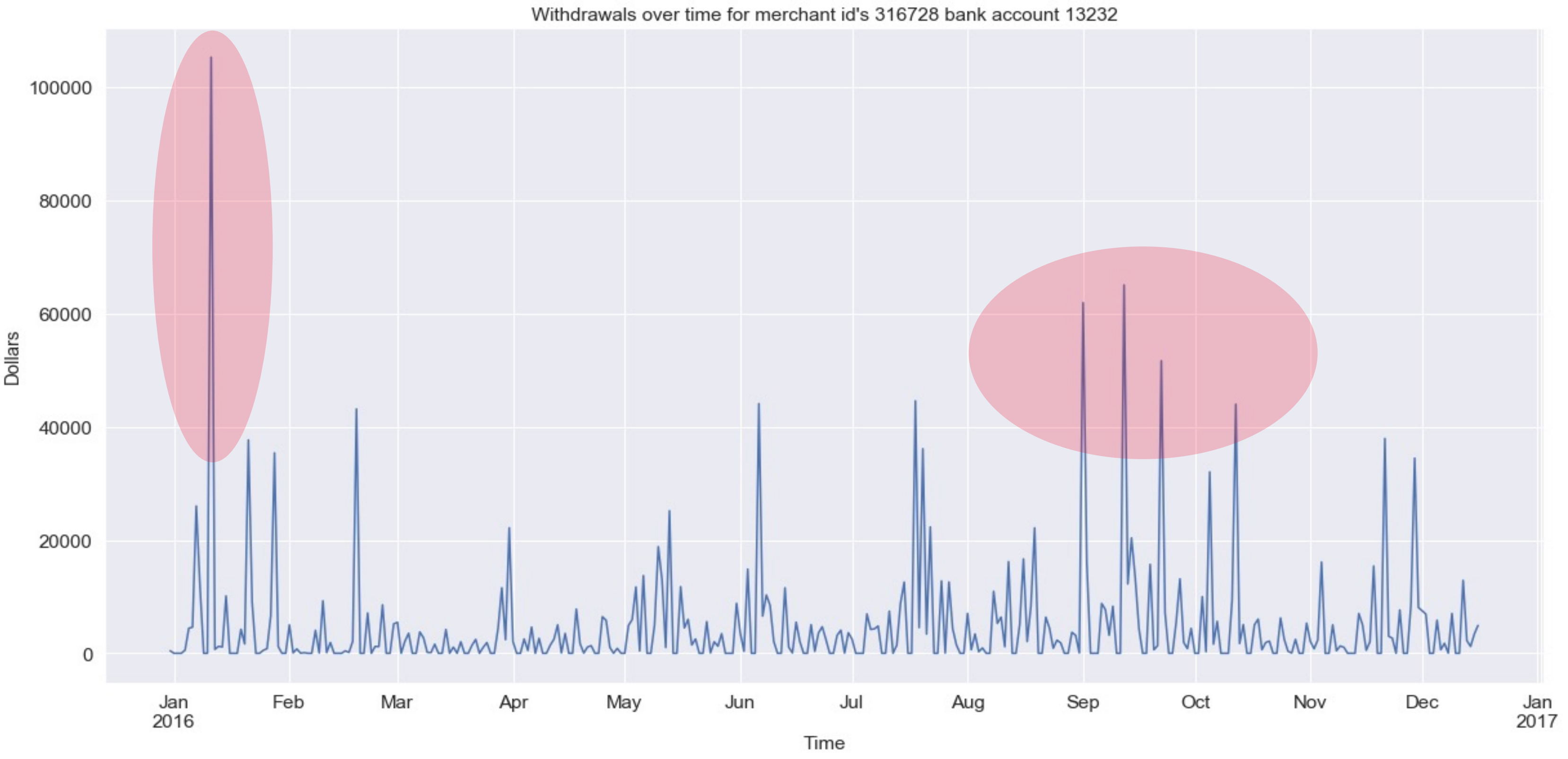
A large deposit happened in November 2016

Daily balance of 318465 merchant’s account id 13419



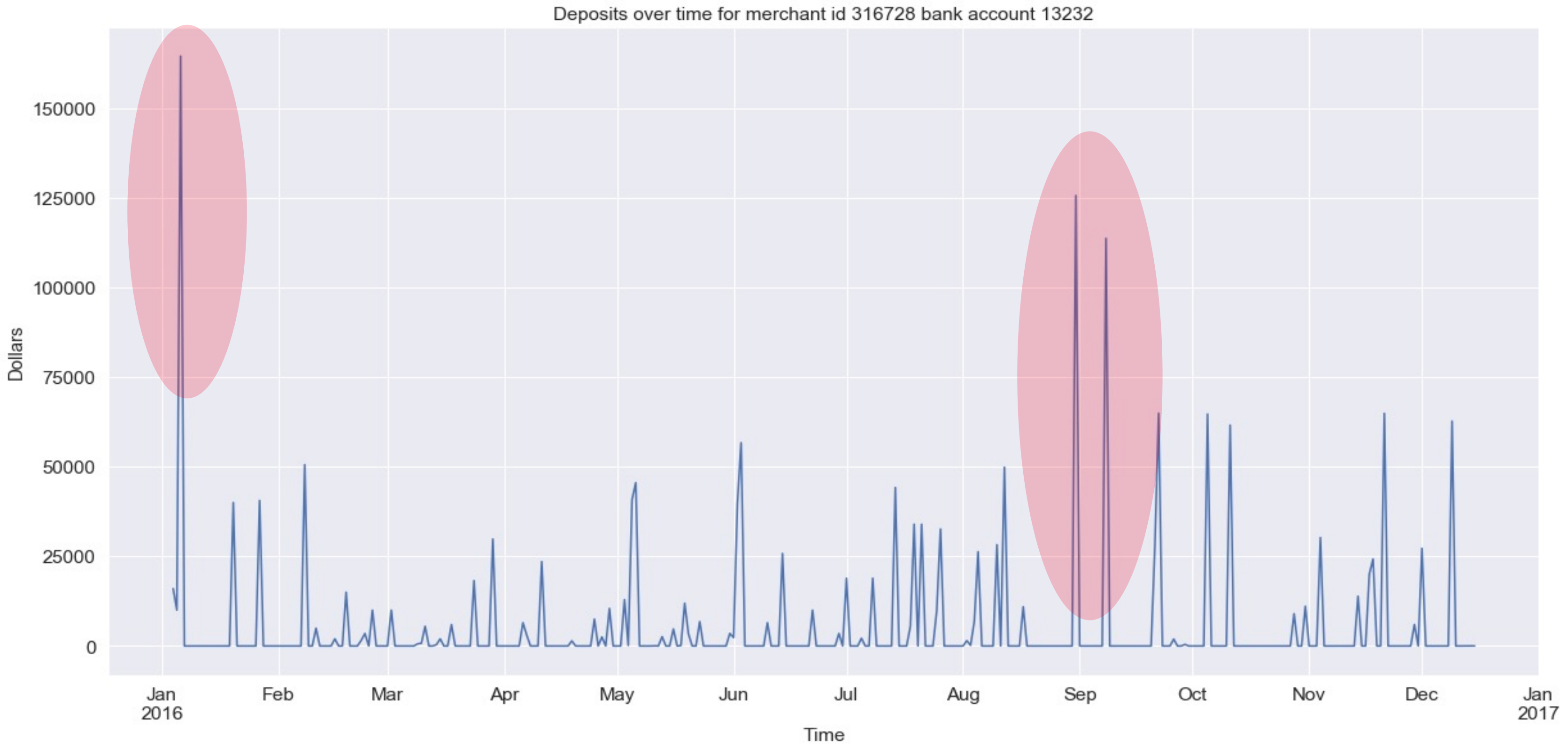
Merchant had negative balance of more than $500 in several months

Debit transaction trend of 316728 merchant’s account id 13232



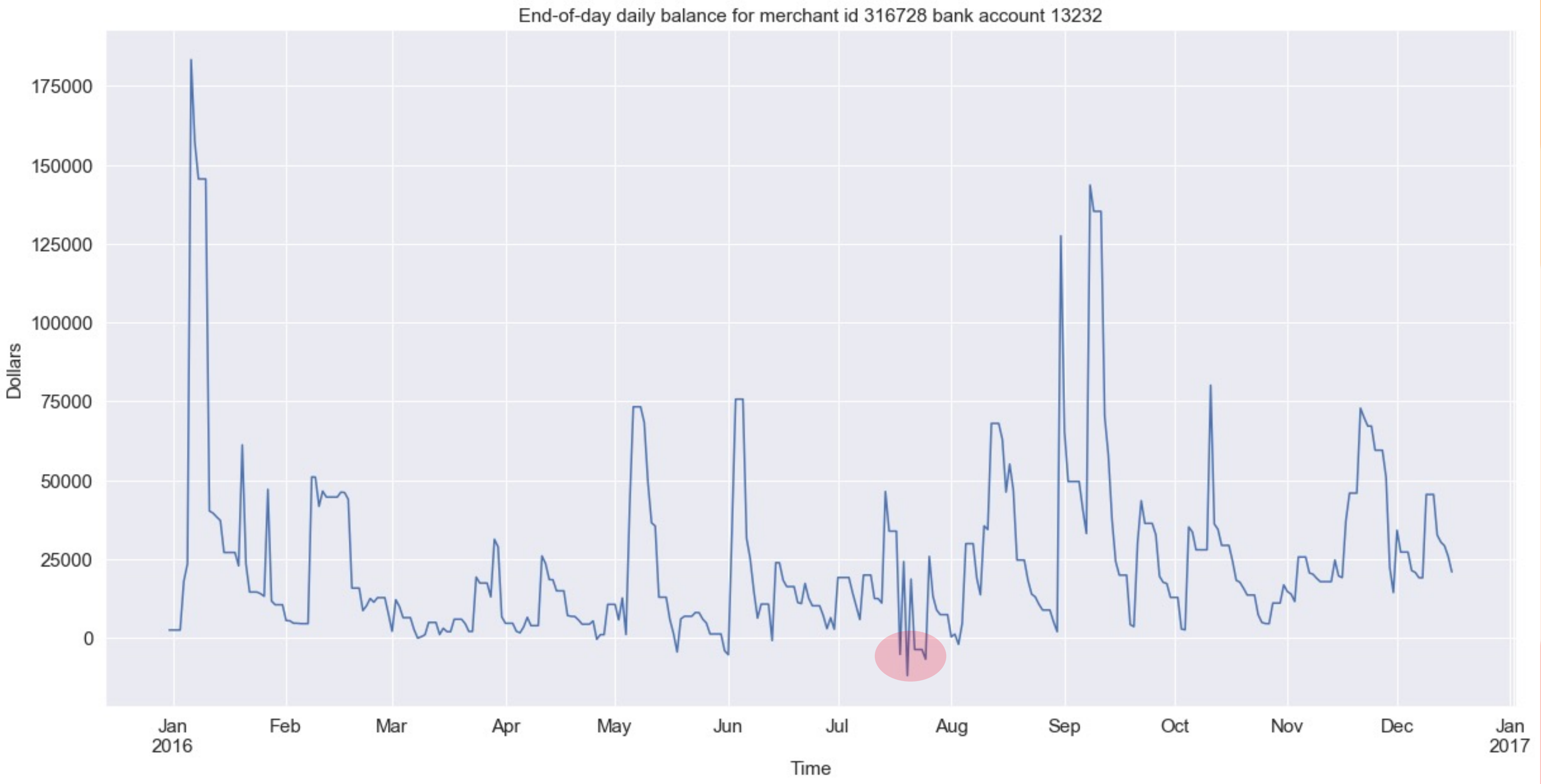
Merchant has made some large withdrawals over time

Credit transaction trend of 316728 merchant’s account id 13232



Larger deposits made into account frequently

Daily balance of 316728 merchant’s account id 13232



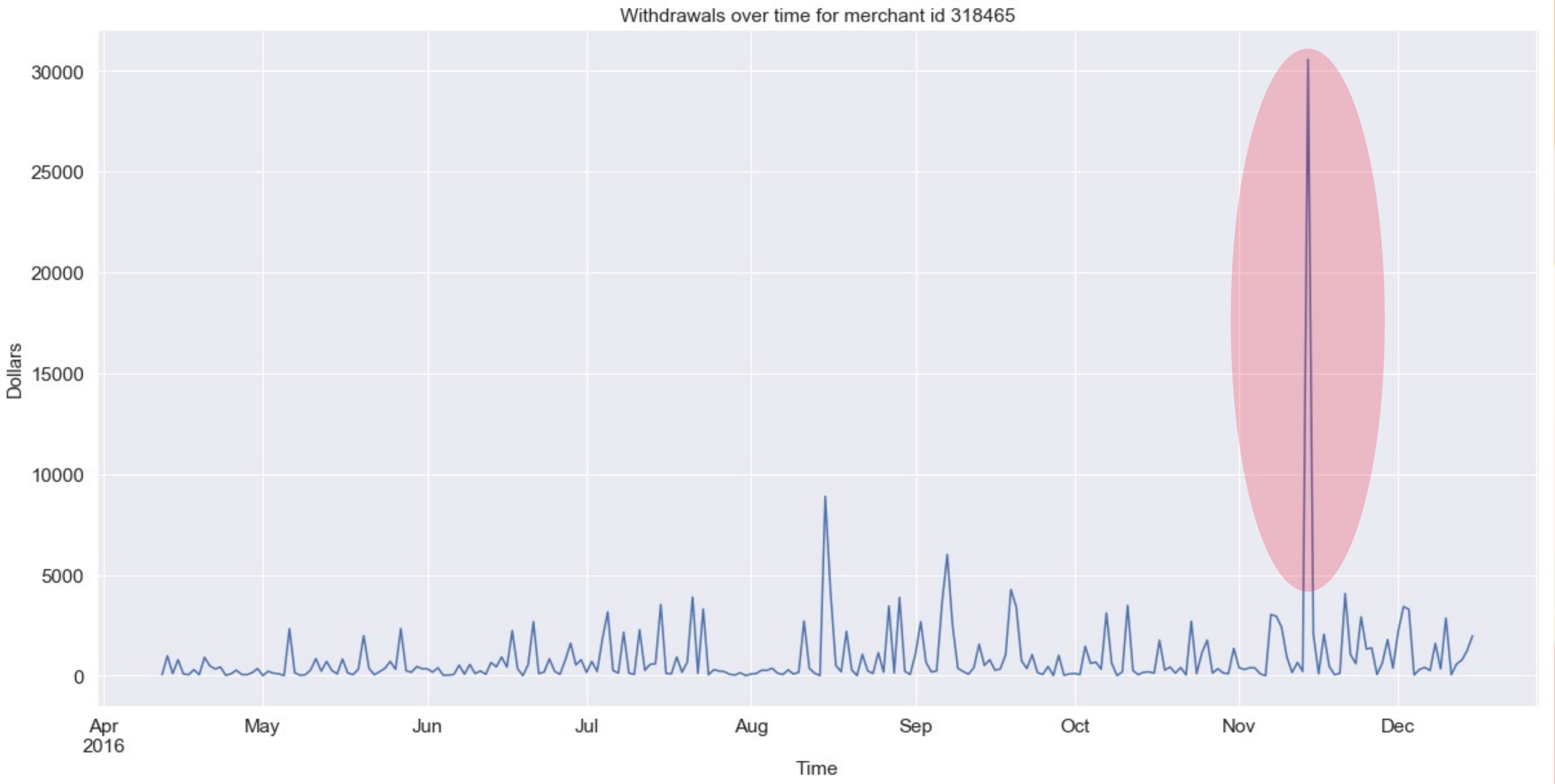
Merchant was able to keep positive daily balance for most periods except few

4.

Time series analysis

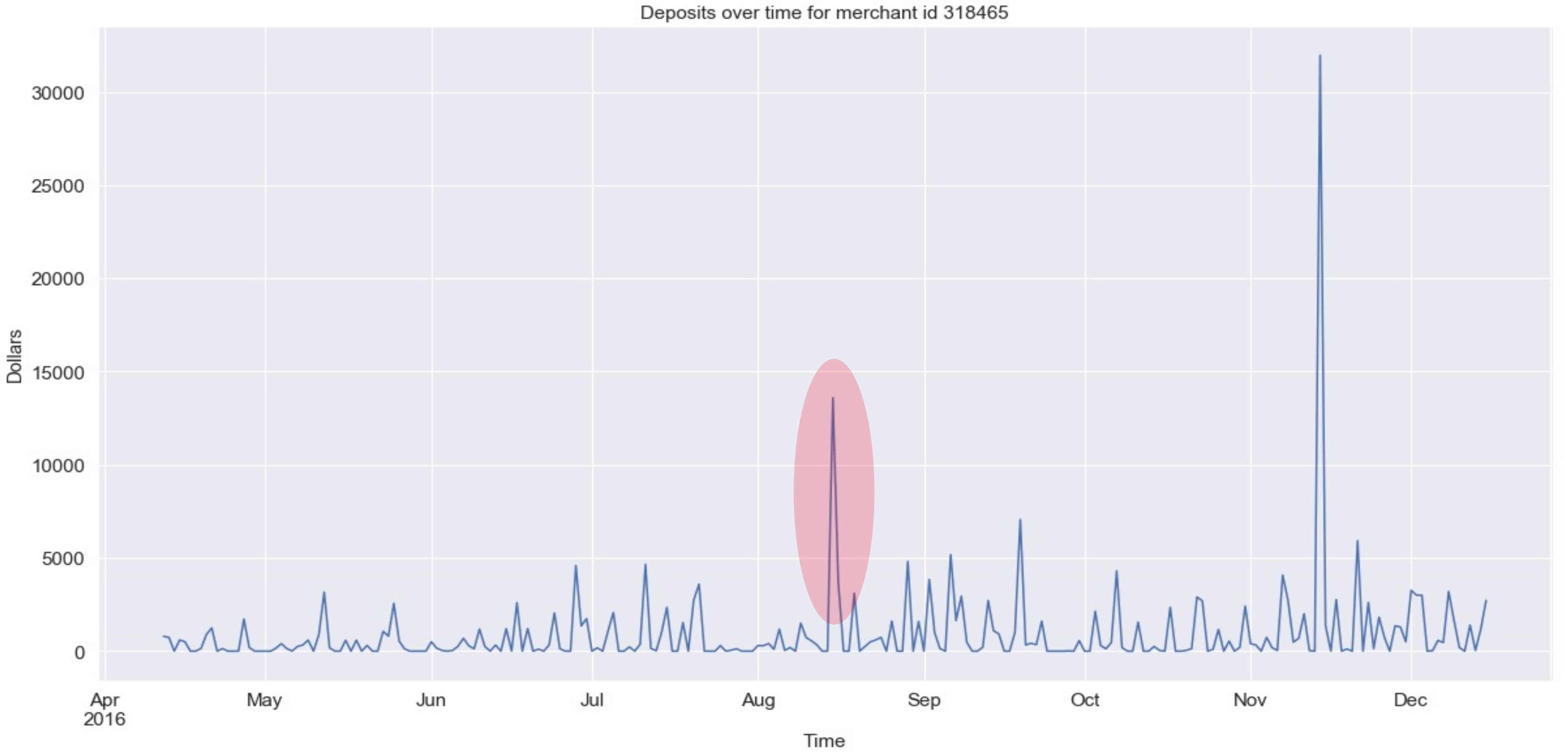
Given Lead ID over all its Bank Account IDs

Debit transaction trend of 318465 aggregated over all bank accounts



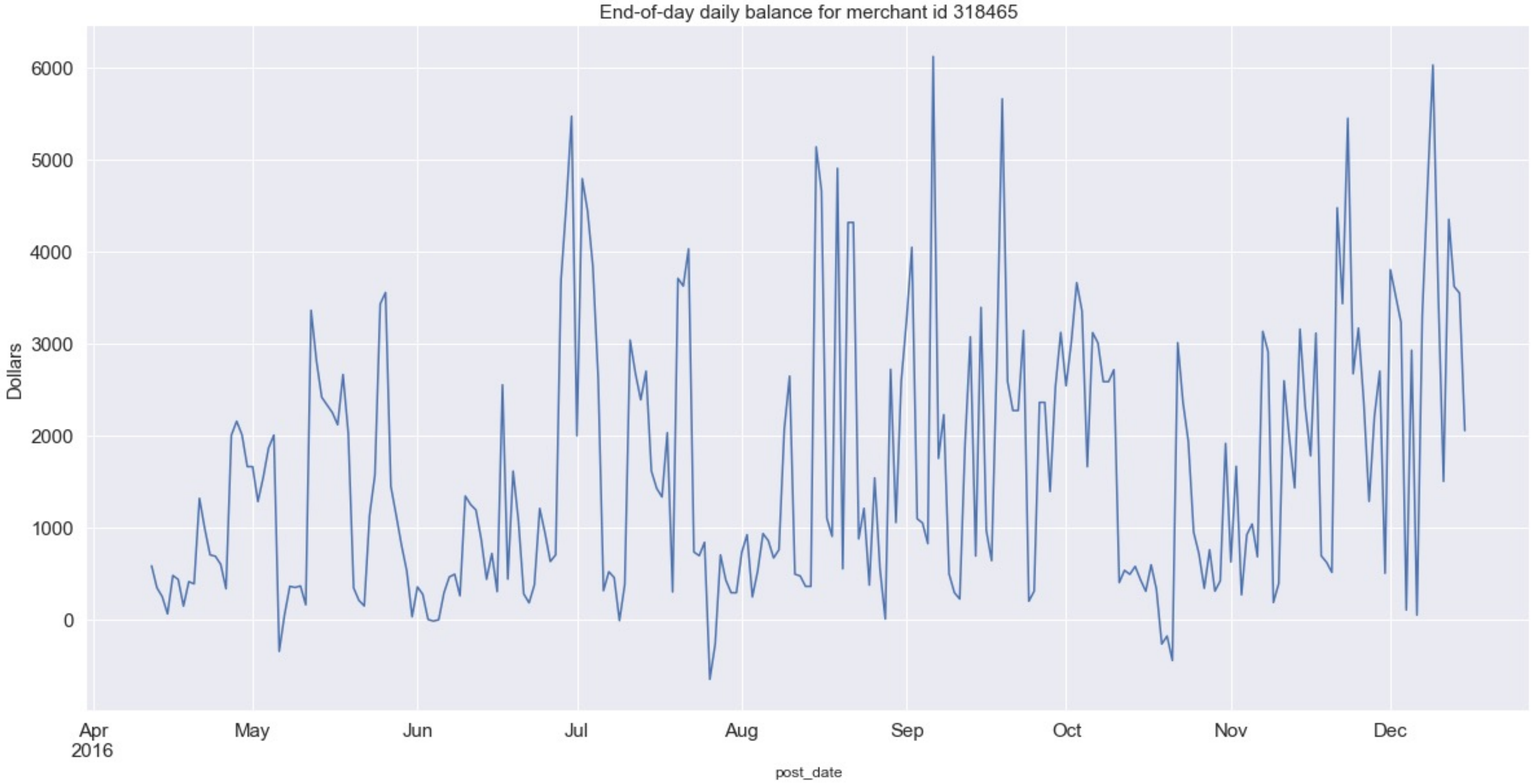
Similar trend as previously presented

Credit transaction trend of 318465 aggregated over all bank accounts



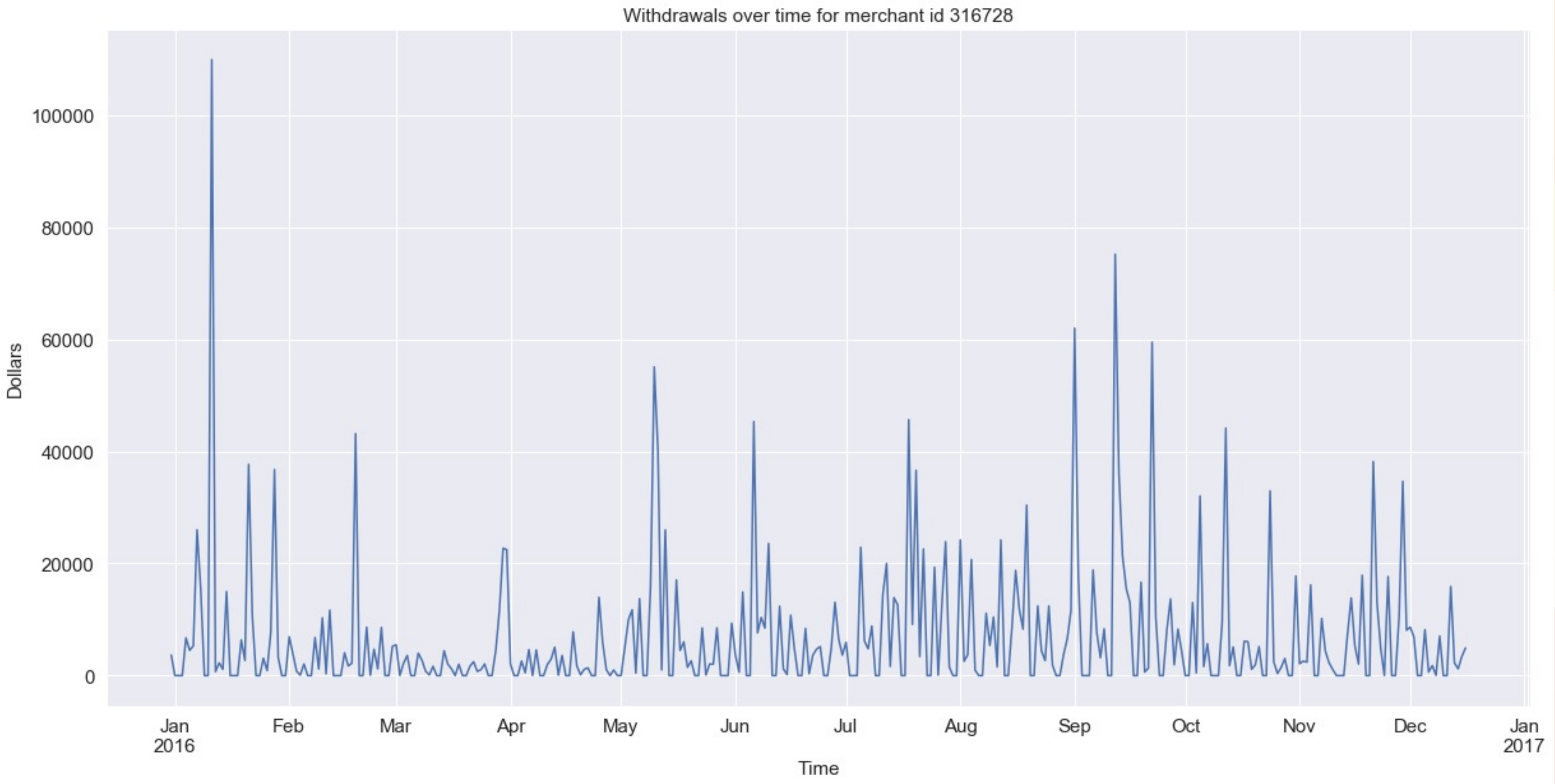
Another bigger credit over all bank account compared to single account

Daily balance of merchant 318465 aggregated over all bank accounts



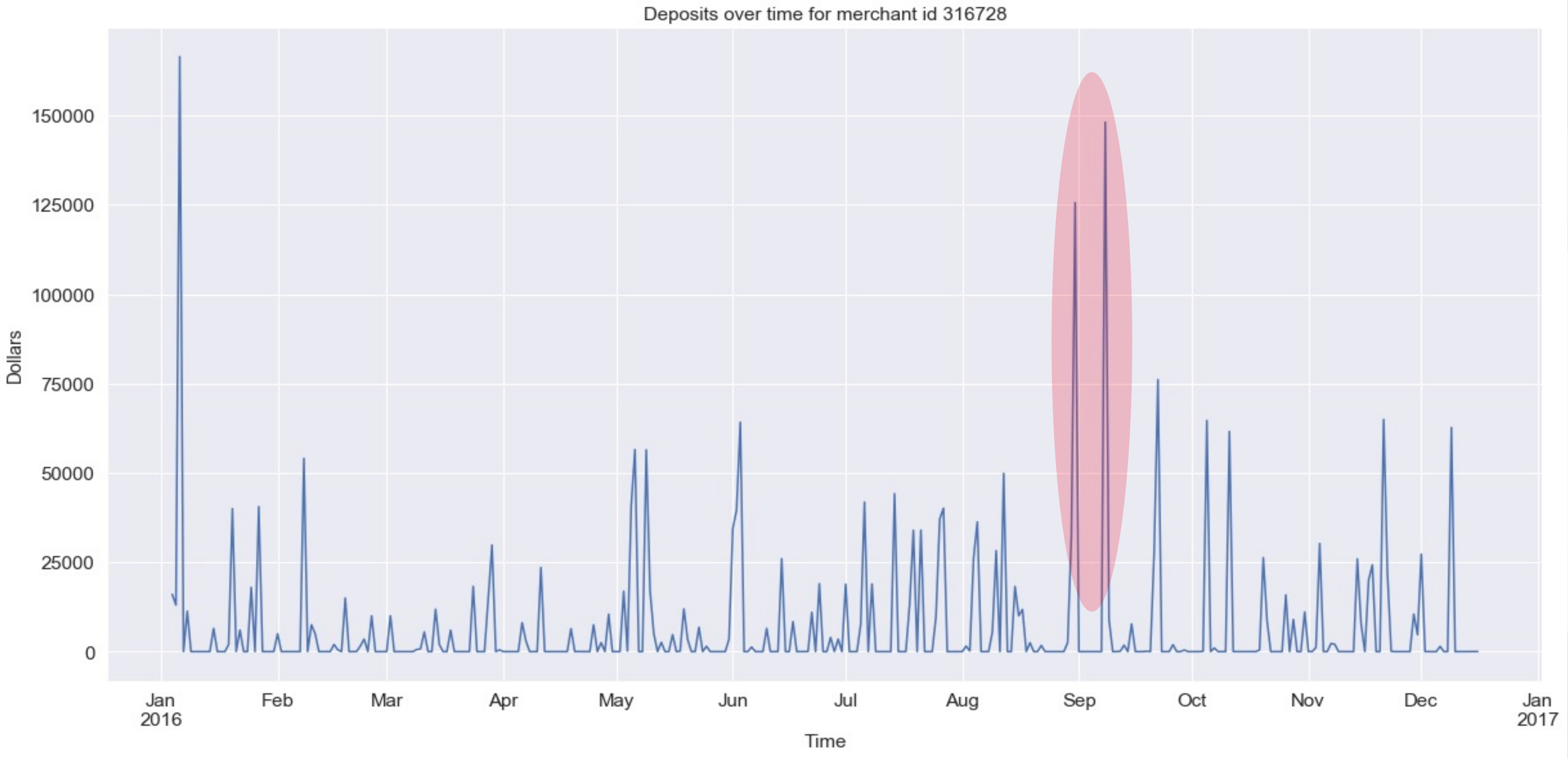
Overall had fewer negative daily balance compared to single bank account

Debit transaction trend of 318465 aggregated over all bank accounts



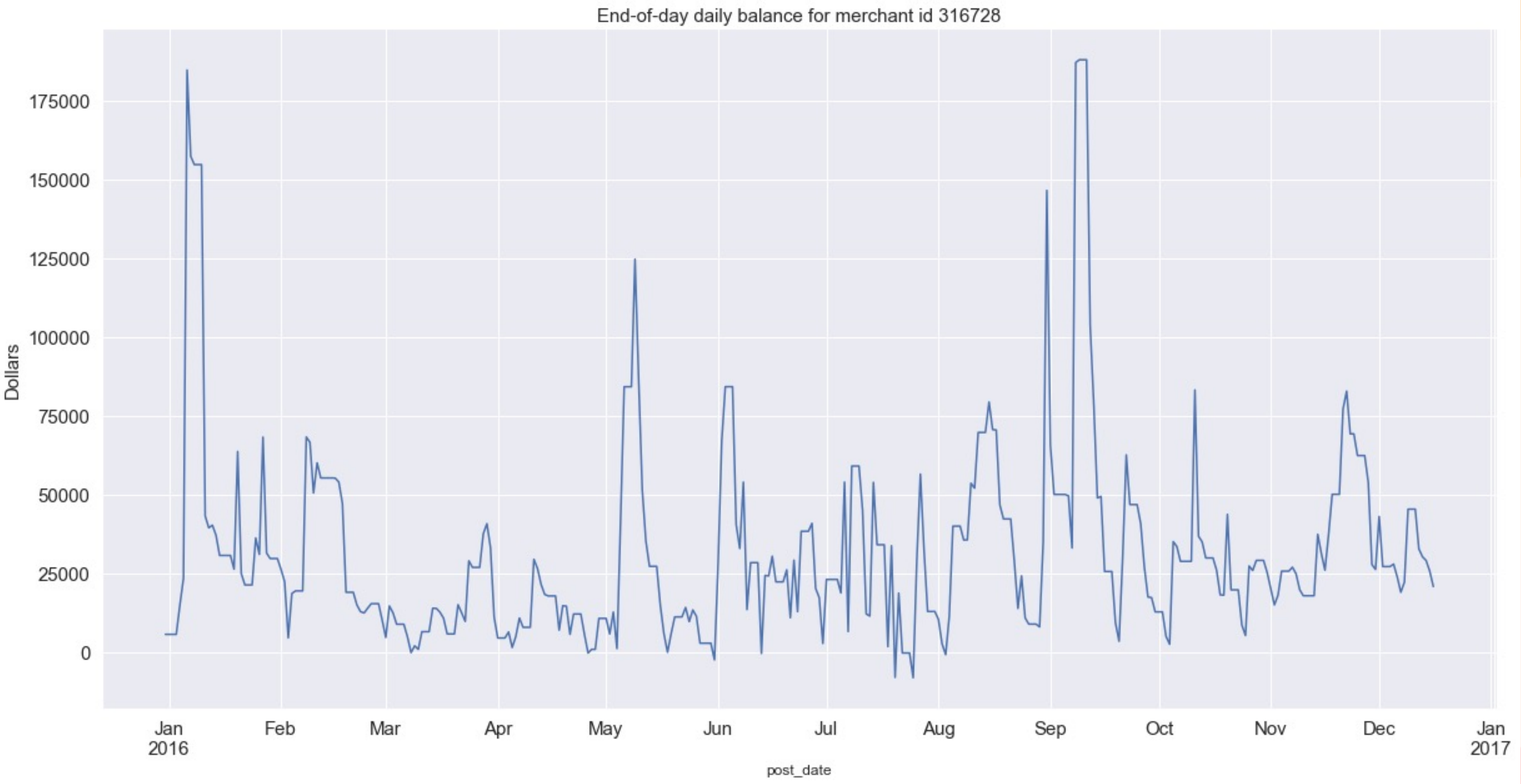
Similar trend as previously presented

Credit transaction trend of 318465 aggregated over all bank accounts



Another bigger credit over all bank account compared to single account

Daily balance of merchant 318465 aggregated over all bank accounts



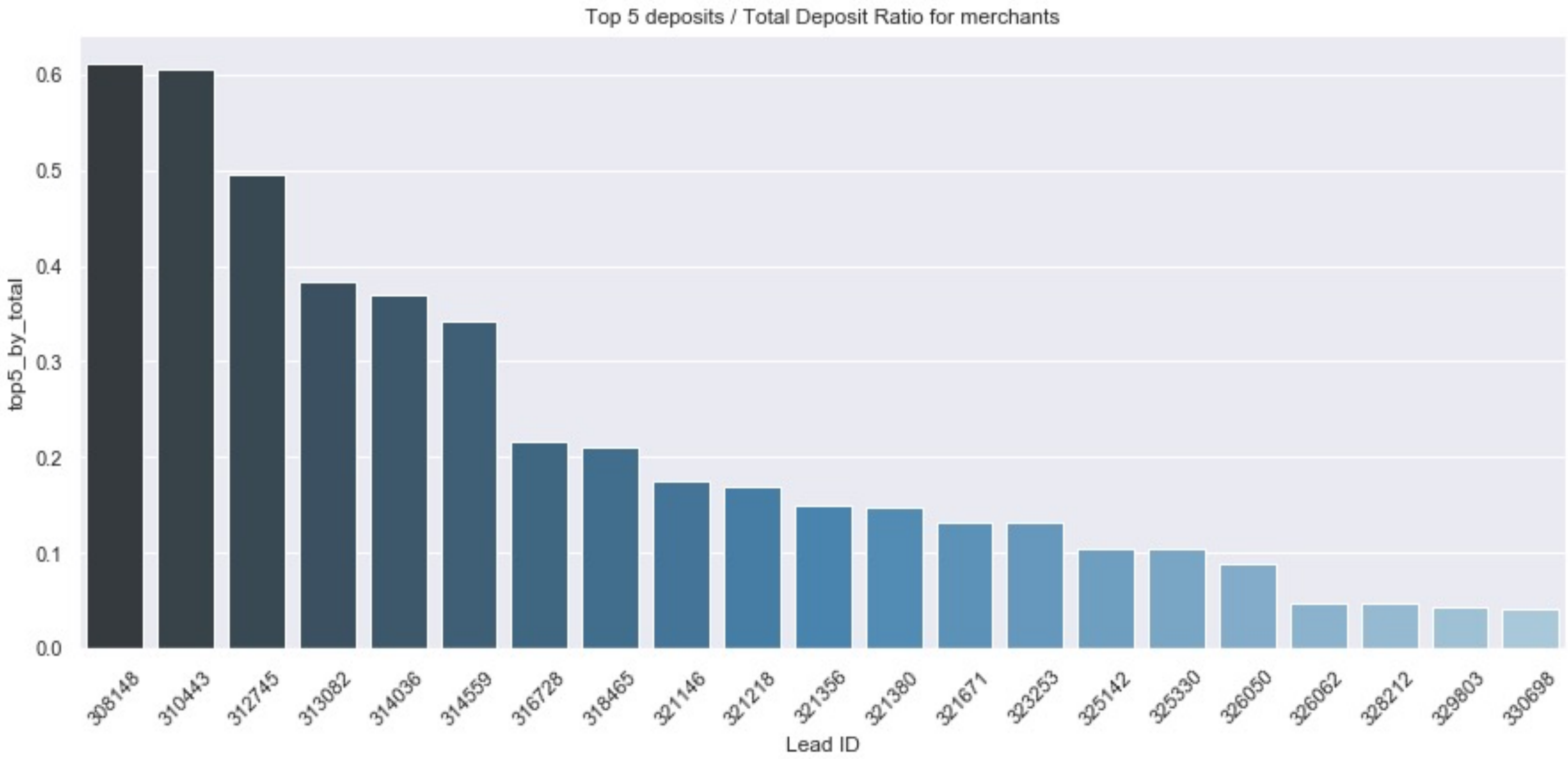
Overall had fewer negative daily balance compared to single bank account

5.

Dimension of cash flow

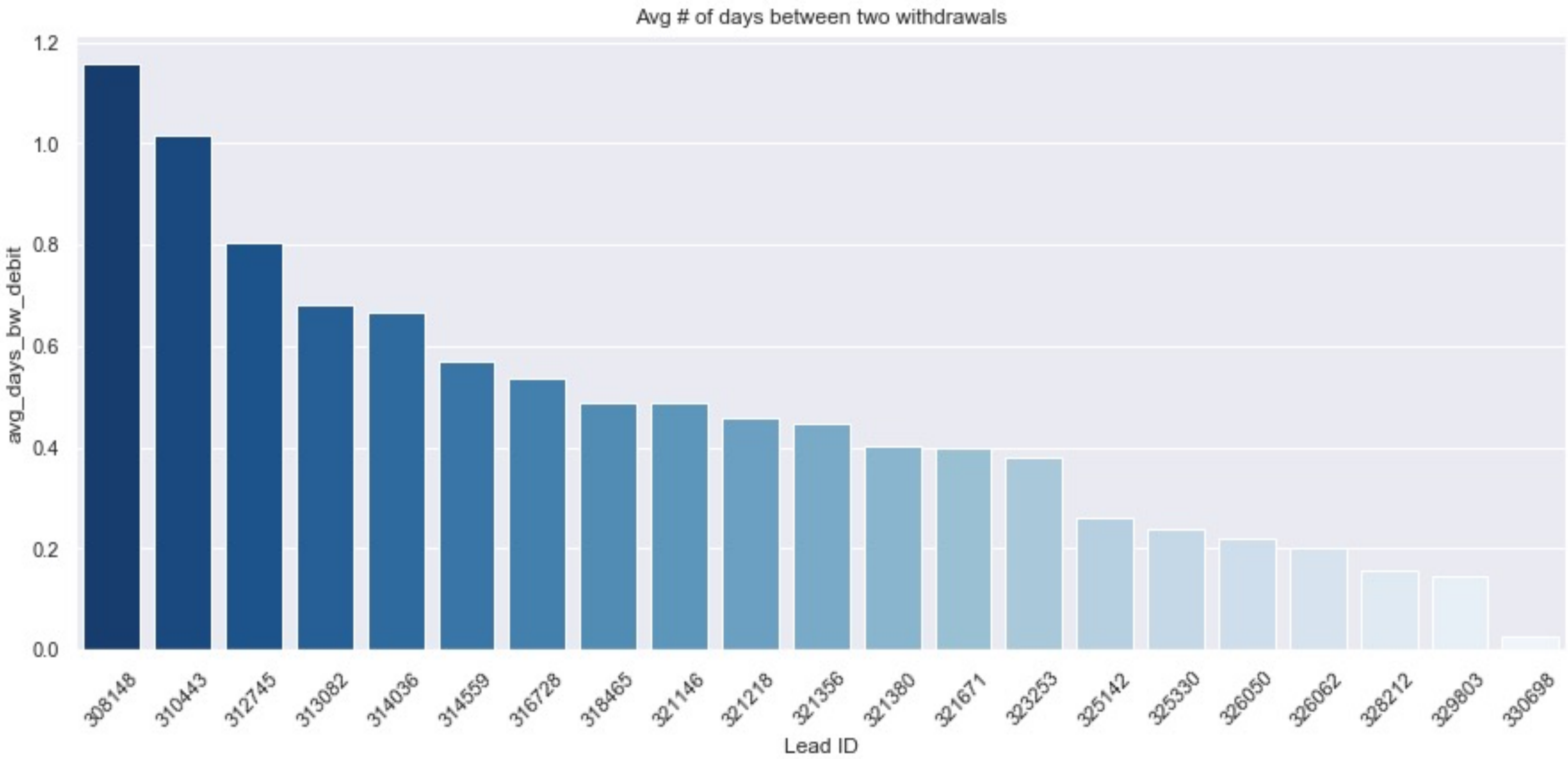
Features that can affect the merchant’s default

Top 5 deposit / Total Deposit



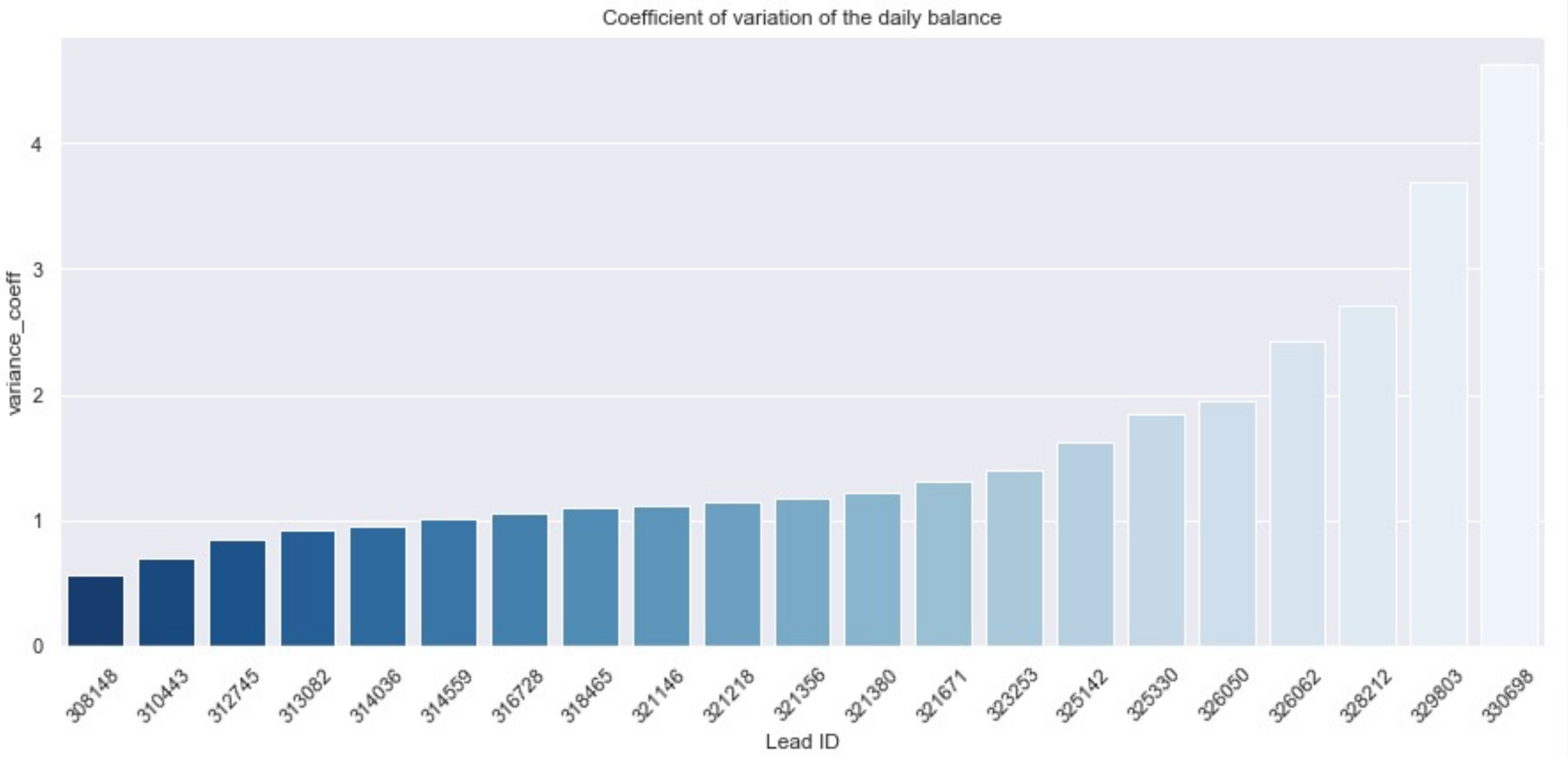
Top 3 merchant ID : 308148, 310443, 312745

Average # of days between two withdrawals



Top 3 merchant ID : 308148, 310443, 312745

Variation coefficient for the daily balance



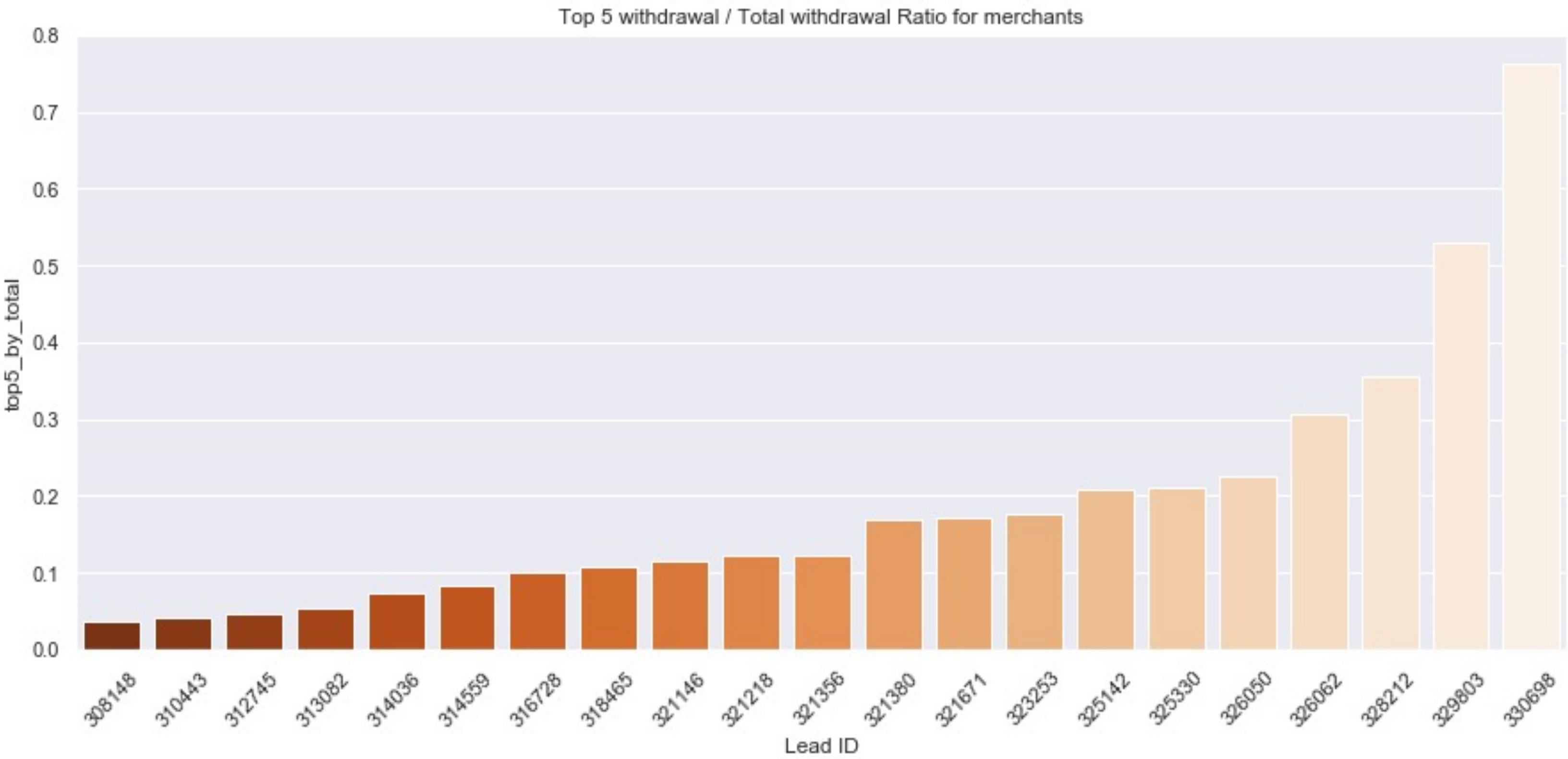
Merchant IDs with least variation: 308148, 310443, 312745

6.

Other Hypothesized Dimensions

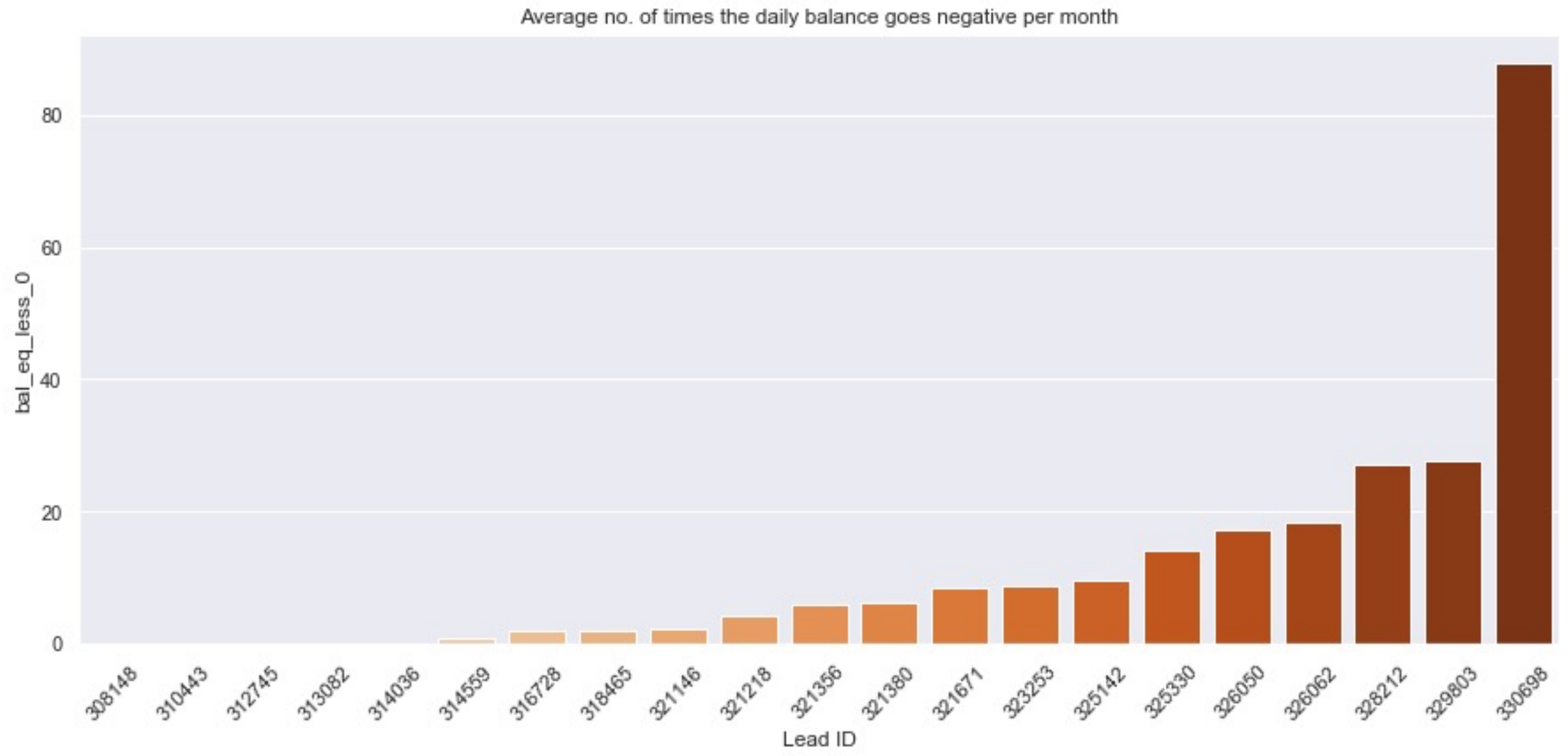
Features that can affect the merchant’s default

Top 5 withdrawal / Total withdrawal



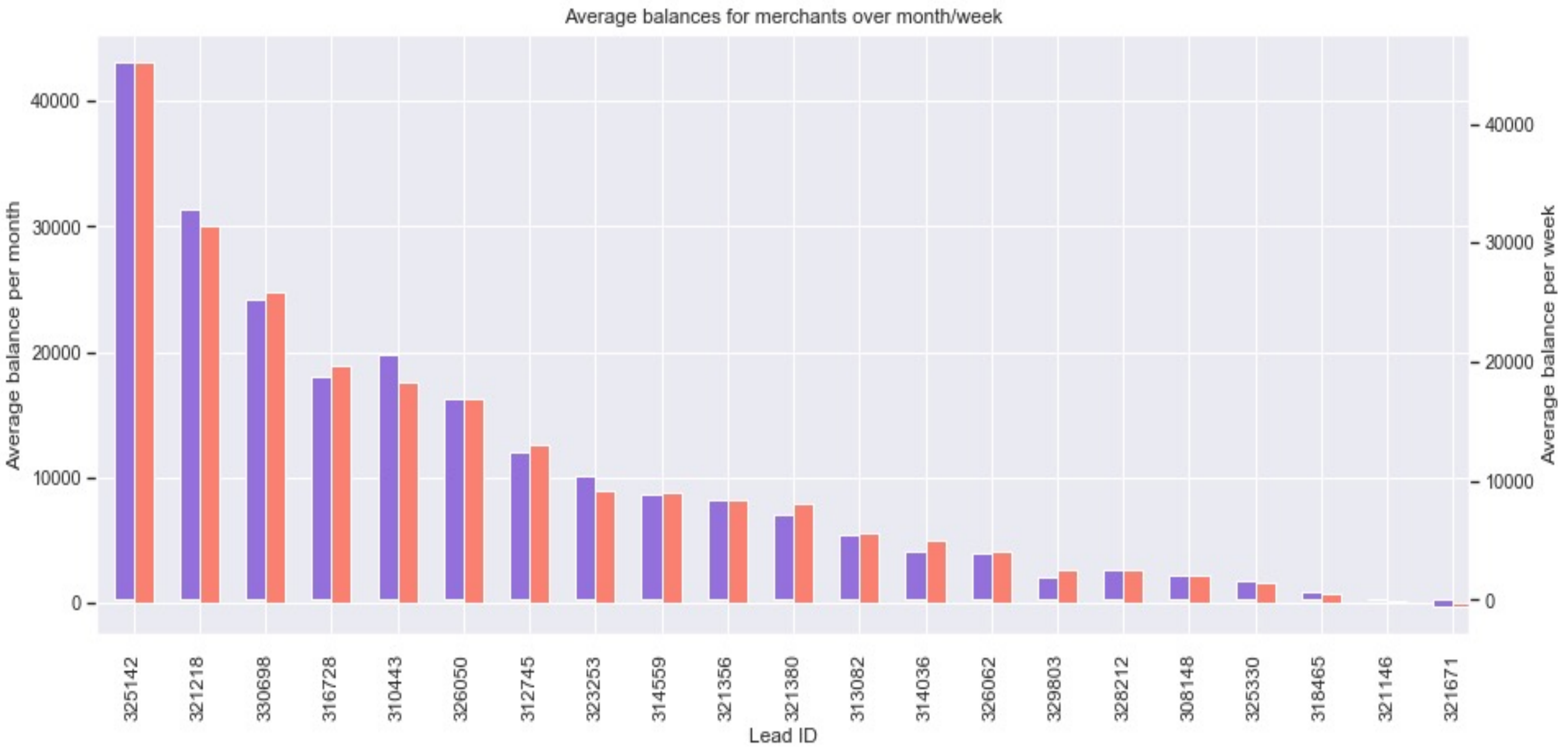
Merchant IDs with least ration: 308148, 310443, 312745

Average # times daily balance becomes non-positive (less than equal to 0)



Merchant IDs who’s balance never go below 0 : 308148, 310443, 312745

Average daily balances over month and week



Merchant IDs with highest balances: 305142, 321218, 330698

7.

Clustering

Grouping transactions

Steps followed

1. Pre-Processing

- Vectorized description text data
- One-Hot encoding of categorical data
- Dealing with timedate column
- Scaling the data

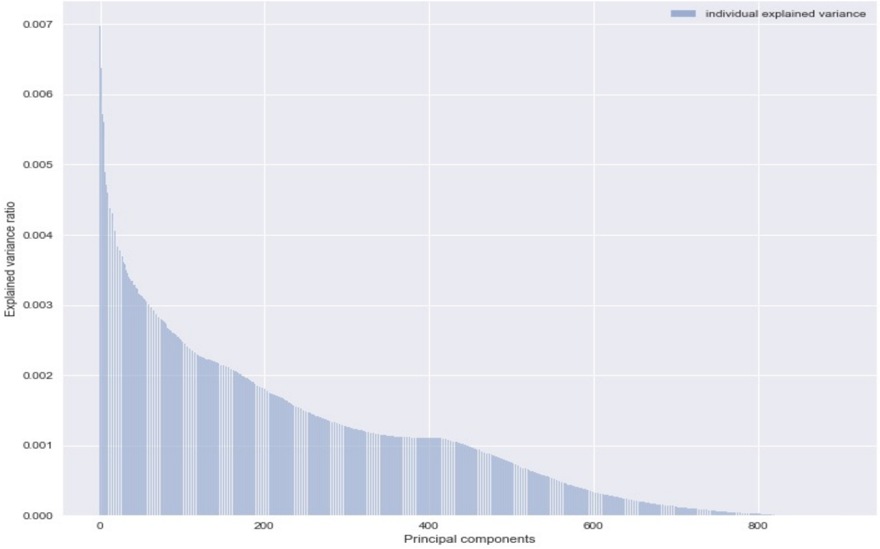
2. PCA

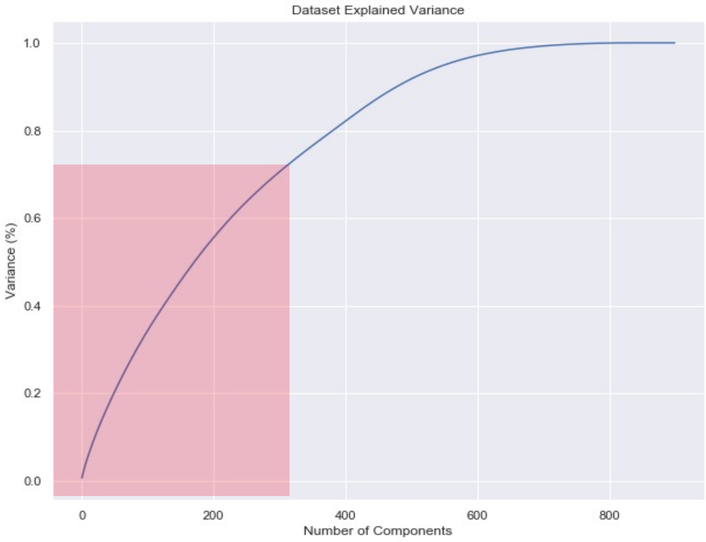
- Performed PCA on data with more than 800 features.
- Selecting top components which explain 75% of variance, which were 300 PCs

3. Clustering

Performed k-Means clustering into 2 clusters

PCA results





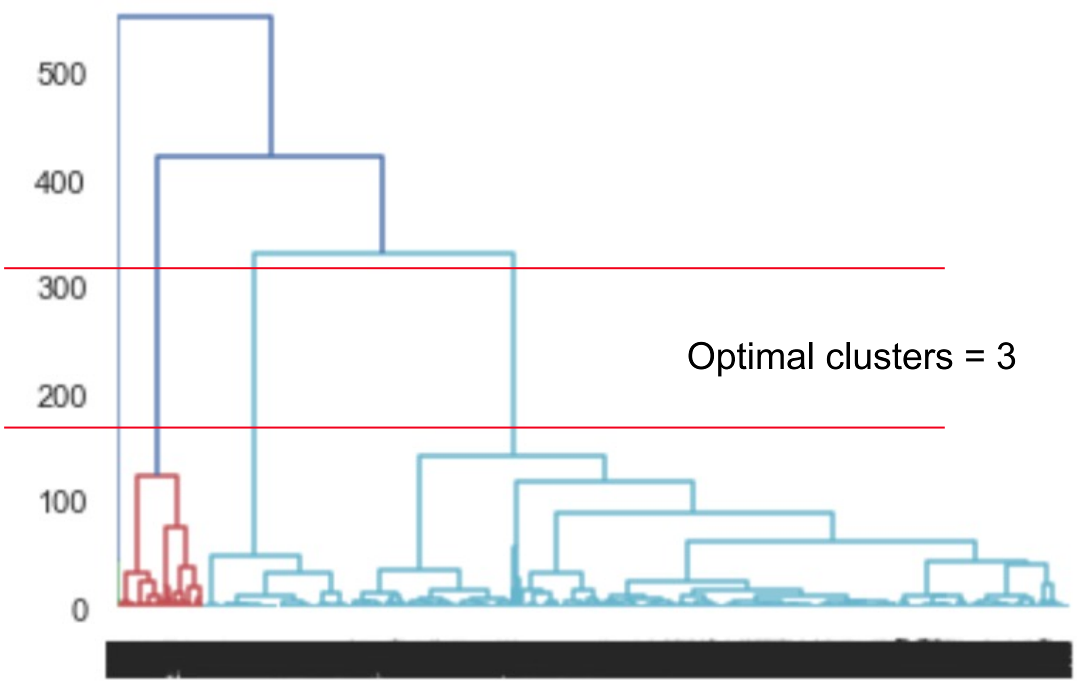
75% of the variance is explained by approx. 300 principal components

Top 5 PCA components out of 300 components

	Component 1	Component 2	Component 3	Component 4	Component 5
Top 5 positive weights	‘wfbt’	'persona'	'trade'	'ending'	'transfer'
	'transfirst’	'doctors'	'bofa'	'indn'	'business'
	'stlmt'	'author'	'world'	'worldpay'	'checking'
	'bkcd'	'bankcard'	'svcs'	'trade'	'datta'
	'breakfast'	'newlogic'	'group'	'bofa'	'online'
Lowest 5 negative weights	'Finance and Insurance'	'transaction_type _code'	'Finance and Insurance'	'transfer'	'Retail Trade’
	'worldpay'	'purchase'	'worldpay'	'checking'	'persona'
	'famil'	'Retail Trade'	'ending’	'business'	'doctors'
	'crystal’	"card'	'indn'	'datta'	'bankcard'
	'lake'	'check'	'waus'	'Other Services’	'author'

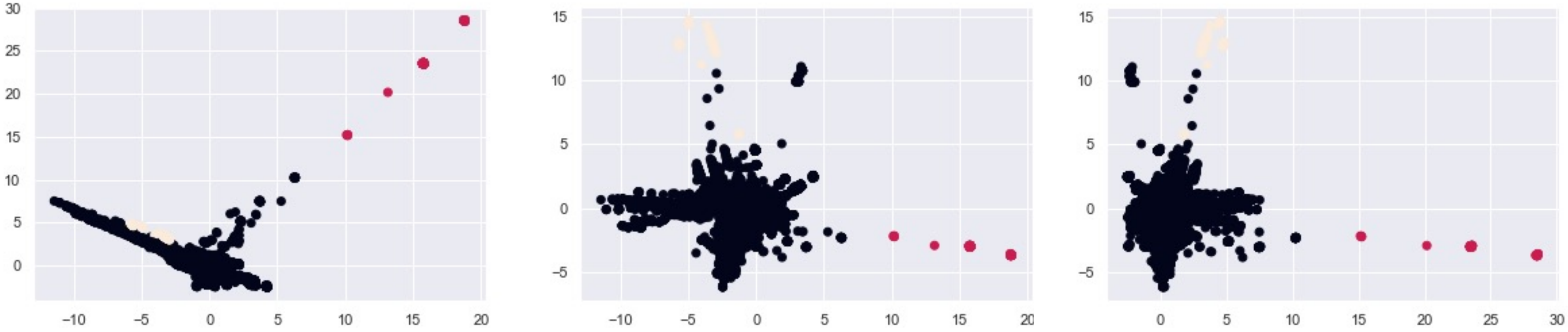
These features have the highest weight values (5 highest positive + 5 lowest negative values)

Dendrogram for finding optimum clusters

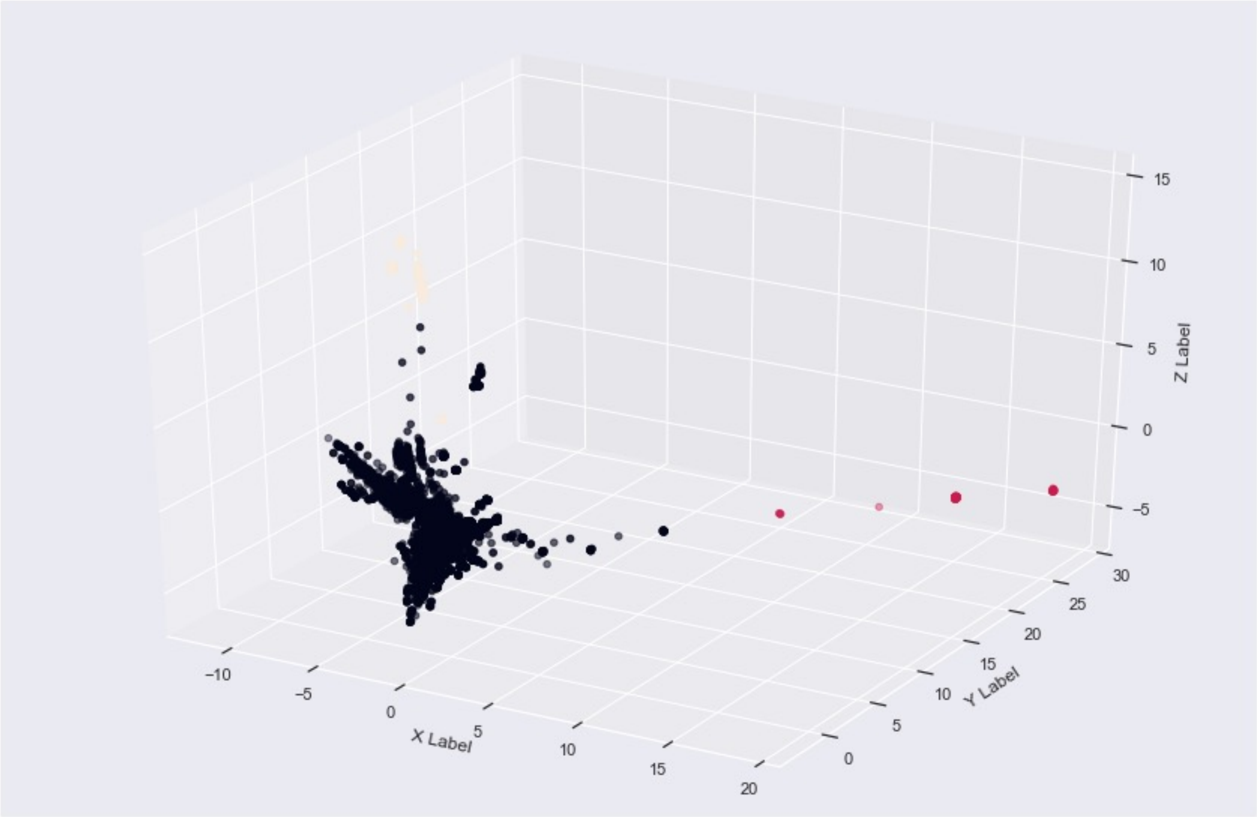


Optimal clusters = 3

K-Means clustering (k = 2)

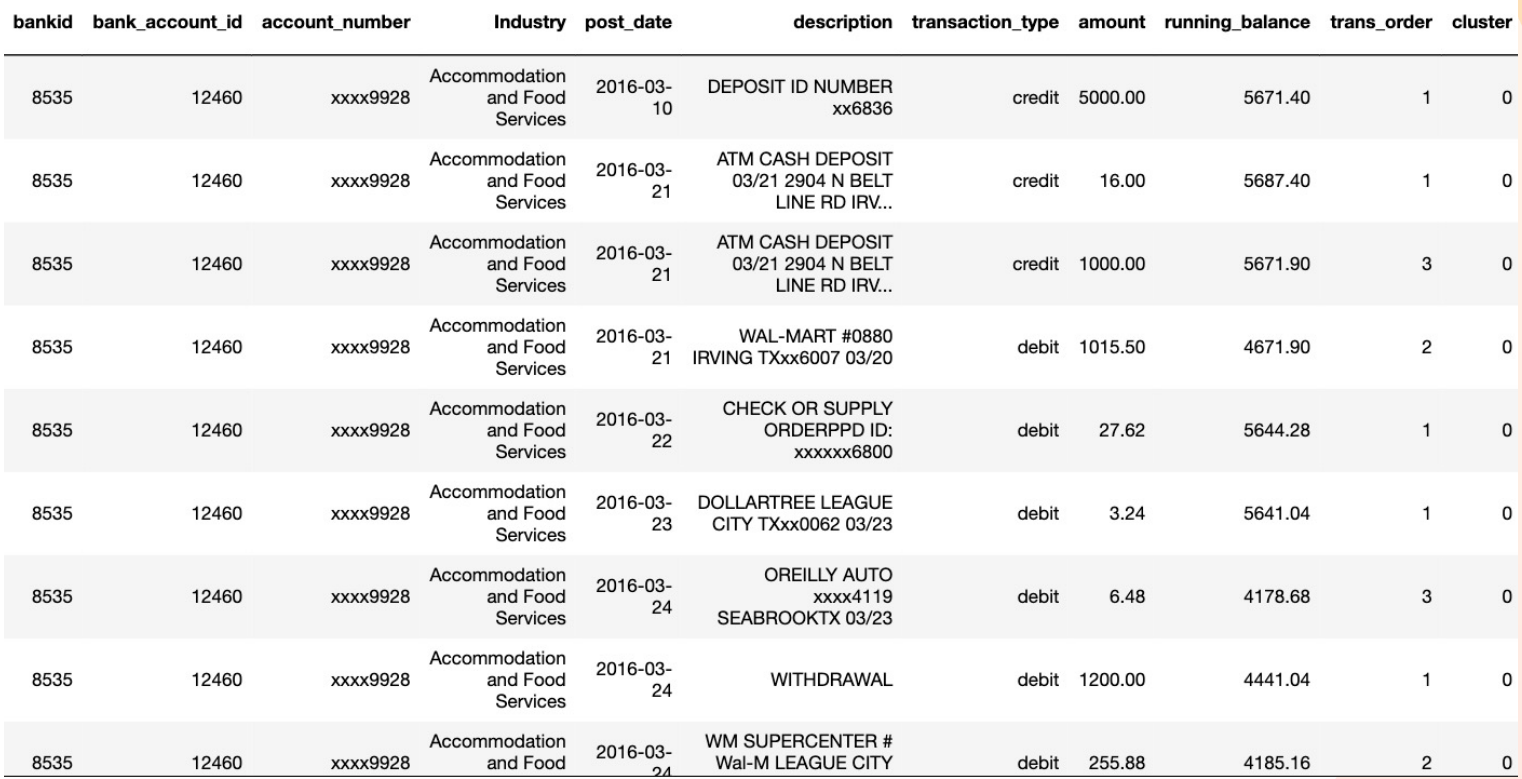


Agglomerative clustering (cluster = 3)



Some clusters for individual merchants

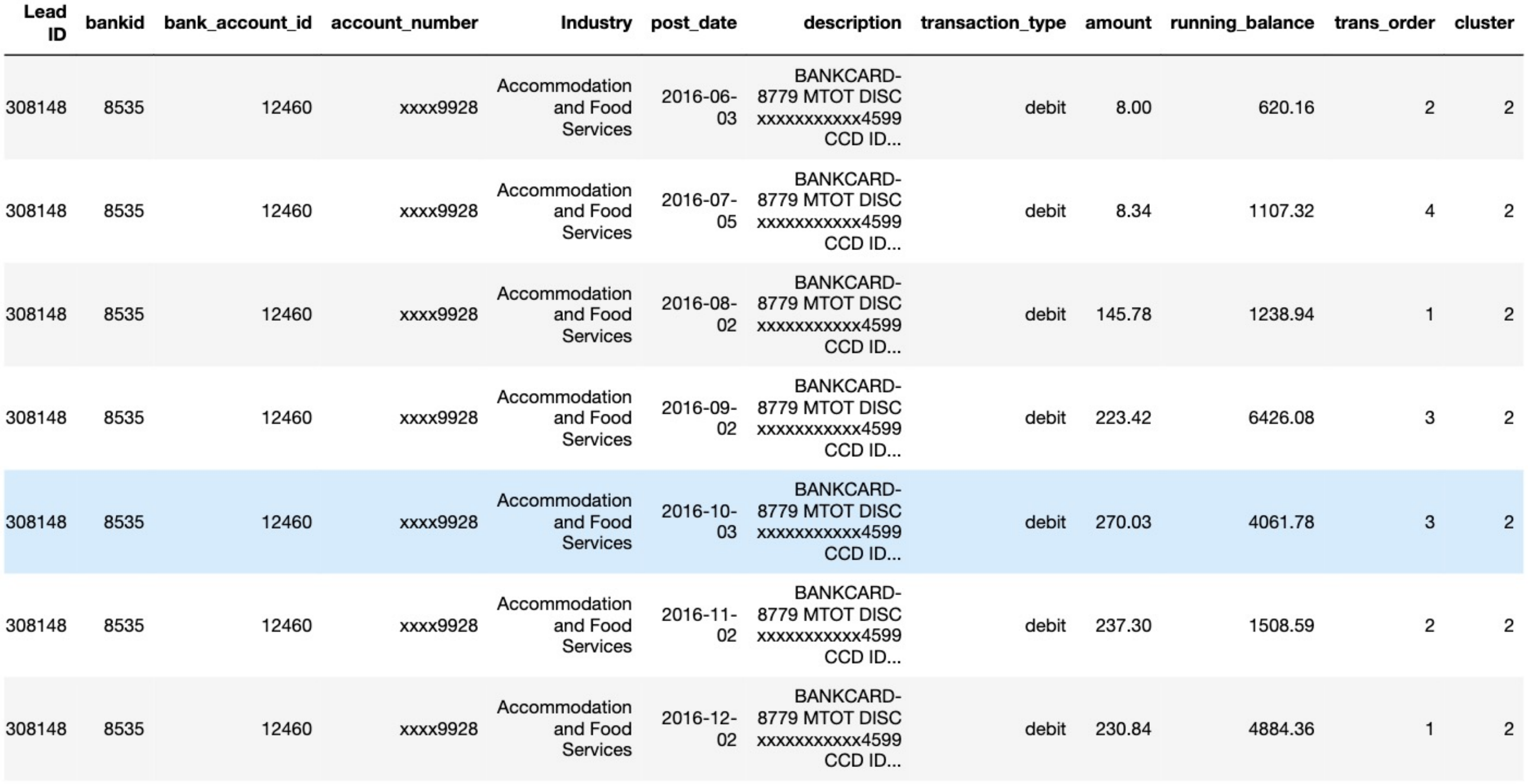
Lead ID – 308148 (Cluster 0)



bankid	bank_account_id	account_number	Industry	post_date	description	transaction_type	amount	running_balance	trans_order	cluster
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-10	DEPOSIT ID NUMBER xx6836	credit	5000.00	5671.40	1	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-21	ATM CASH DEPOSIT 03/21 2904 N BELT LINE RD IRV...	credit	16.00	5687.40	1	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-21	ATM CASH DEPOSIT 03/21 2904 N BELT LINE RD IRV...	credit	1000.00	5671.90	3	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-21	WAL-MART #0880 IRVING TXxx6007 03/20	debit	1015.50	4671.90	2	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-22	CHECK OR SUPPLY ORDERPPD ID: xxxxxx6800	debit	27.62	5644.28	1	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-23	DOLLARTREE LEAGUE CITY TXxx0062 03/23	debit	3.24	5641.04	1	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-24	OREILLY AUTO xxxx4119 SEABROOKTX 03/23	debit	6.48	4178.68	3	0
8535	12460	xxxx9928	Accommodation and Food Services	2016-03-24	WITHDRAWAL	debit	1200.00	4441.04	1	0
8535	12460	xxxx9928	Accommodation and Food	2016-03-24	WM SUPERCENTER # Wal-M LEAGUE CITY	debit	255.88	4185.16	2	0

Some clusters for individual merchants

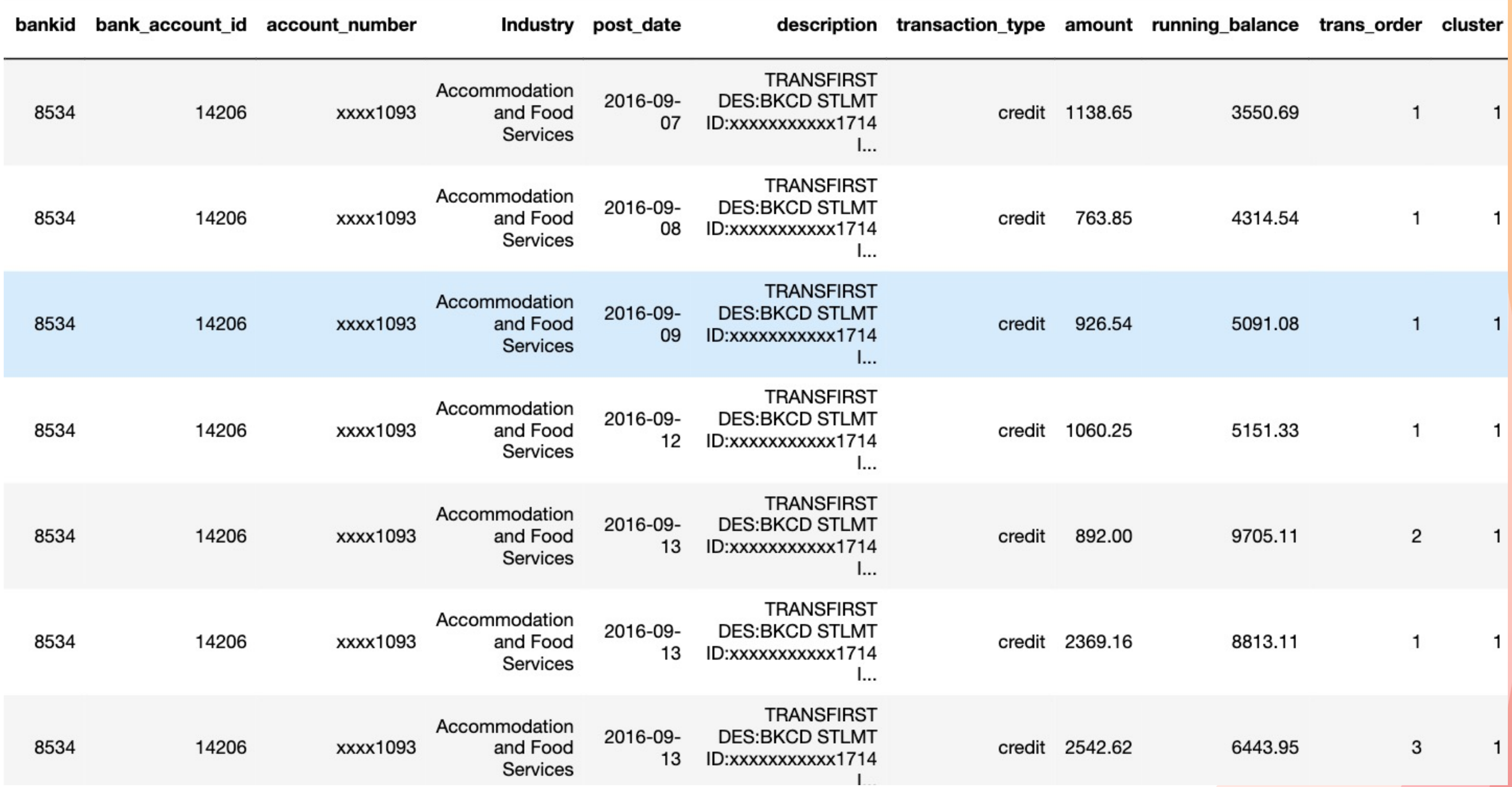
Lead ID – 308148 (Cluster 2)



Lead ID	bankid	bank_account_id	account_number	Industry	post_date	description	transaction_type	amount	running_balance	trans_order	cluster
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-06-03	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	8.00	620.16	2	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-07-05	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	8.34	1107.32	4	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-08-02	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	145.78	1238.94	1	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-09-02	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	223.42	6426.08	3	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-10-03	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	270.03	4061.78	3	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-11-02	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	237.30	1508.59	2	2
308148	8535	12460	xxxx9928	Accommodation and Food Services	2016-12-02	BANKCARD-8779 MTOT DISC xxxxxxxxxxxx4599 CCD ID...	debit	230.84	4884.36	1	2

Some clusters for individual merchants

Lead ID – 326050 (Cluster 1)



bankid	bank_account_id	account_number	Industry	post_date	description	transaction_type	amount	running_balance	trans_order	cluster
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-07	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	1138.65	3550.69	1	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-08	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	763.85	4314.54	1	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-09	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	926.54	5091.08	1	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-12	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	1060.25	5151.33	1	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-13	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	892.00	9705.11	2	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-13	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	2369.16	8813.11	1	1
8534	14206	xxxx1093	Accommodation and Food Services	2016-09-13	TRANSFIRST DES:BKCD STLMT ID:xxxxxxxxxxx1714 I...	credit	2542.62	6443.95	3	1

Thanks!