CAPSTONE 2

Elo Merchant Category Recommendation

Helps understand customer loyalty

INDEX

Capstone 2 Proposal

- **Problem Statement:**

Elo , one of the largest payment brands in Brazil, has built partnerships with merchants in order to offer promotions or discounts to cardholders. But do these promotions work for either the consumer or the merchant? Do customers enjoy their experience? Do merchants see repeat business? Personalization is key.

Elo has built machine learning models to understand the most important aspects and preferences in their customers' lifecycle, from food to shopping. But so far none of them is specifically tailored for an individual or profile.

In this competition, it is expected to develop algorithms to identify and serve the most relevant opportunities to individuals, by uncovering signal in customer loyalty. Your input will improve customers' lives and help Elo reduce unwanted campaigns, to create the right experience for customers.Performance evaluation algorithm is expected to be RMSE.

- **What data are you using? How will you acquire the data?**

The data is available on Kaggle.The link is as below:

https://www.kaggle.com/c/elo-merchant-category-recommendation/data

- **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

  I will try to solve this problem using regression analysis by various machine learning algorithms to predict customer loyalty as target.

- **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

Deliverables would include ,detailed Exploratory data analysis,Inferential statistics and Building machine learning model.

All of the above will be part of github repository ,along with slides and blog post.

DATA WRANGLING

The dataset contains various files named below:

1.train.csv

2.test.csv

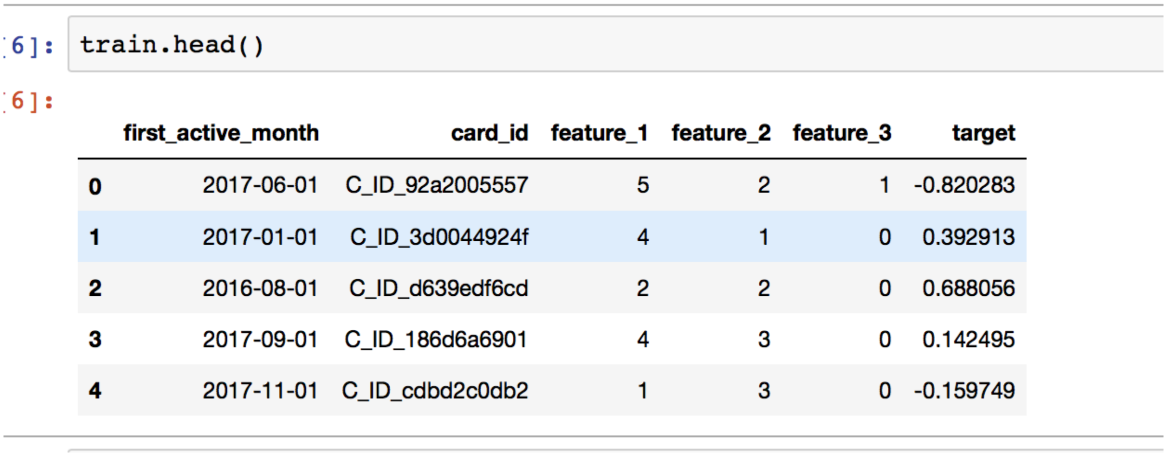
3.merchants.csv

4.historical_transactions.csv

5.new_merchant_transactions.csv

Lets deep dive into various csv files.

If we observe train.csv



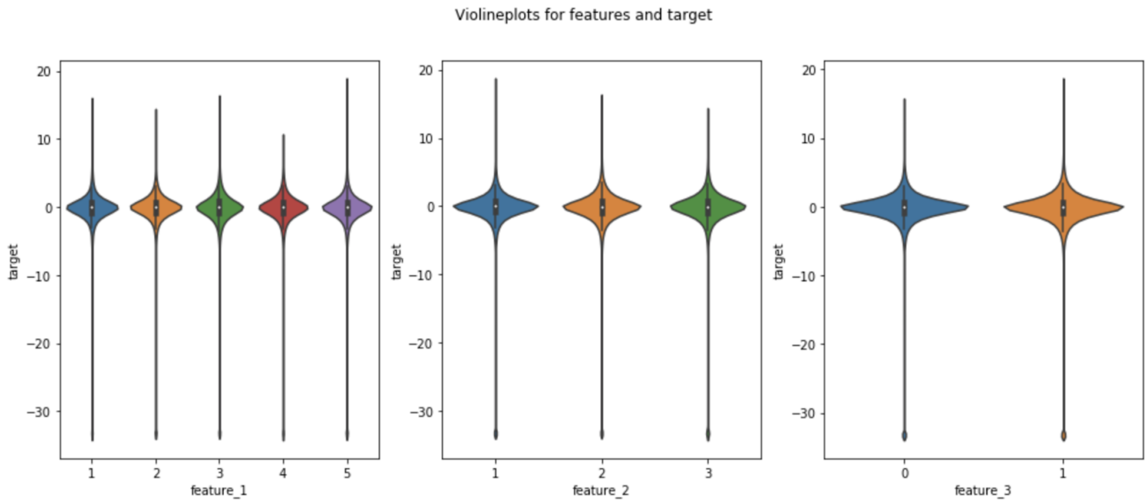
```
6]: train.head()
```

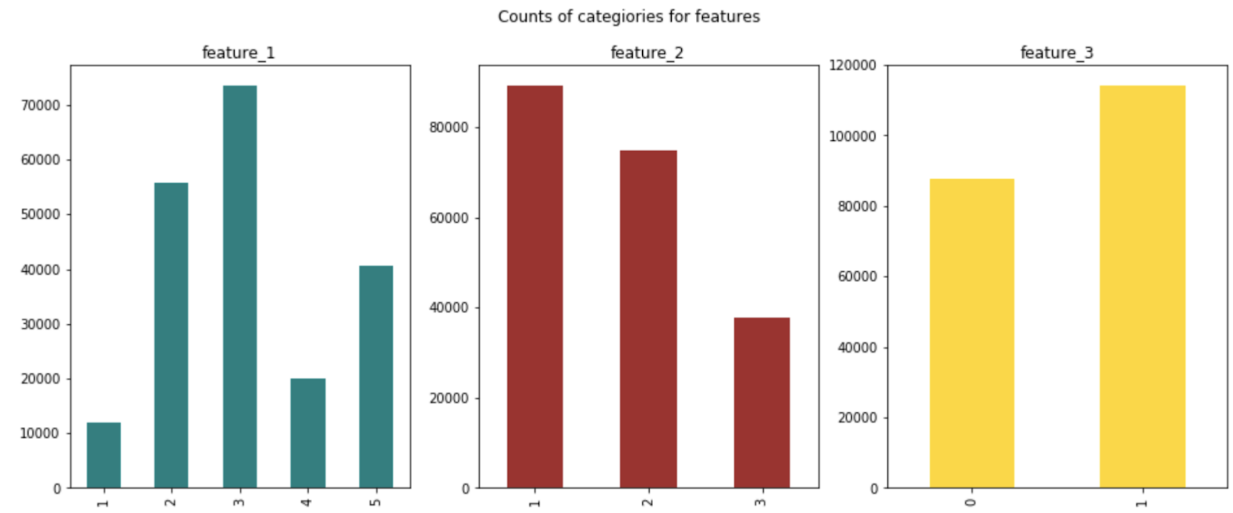
```
6]:
```

	first_active_month	card_id	feature_1	feature_2	feature_3	target
0	2017-06-01	C_ID_92a2005557	5	2	1	-0.820283
1	2017-01-01	C_ID_3d0044924f	4	1	0	0.392913
2	2016-08-01	C_ID_d639edf6cd	2	2	0	0.688056
3	2017-09-01	C_ID_186d6a6901	4	3	0	0.142495
4	2017-11-01	C_ID_cdbd2c0db2	1	3	0	-0.159749

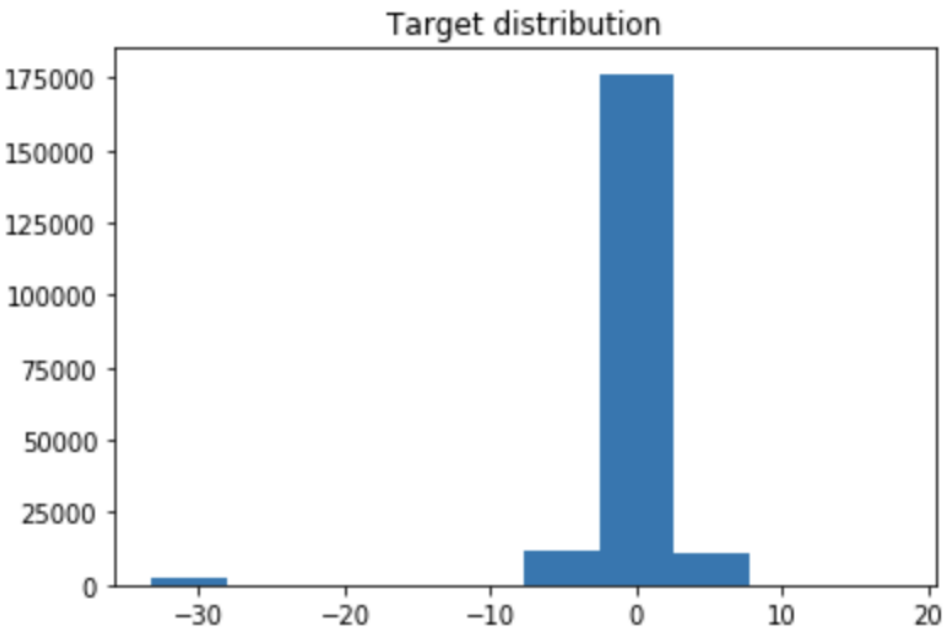
Feature_1,feature_2 and feature_3 are anonymised features ,target is the loyalty score.

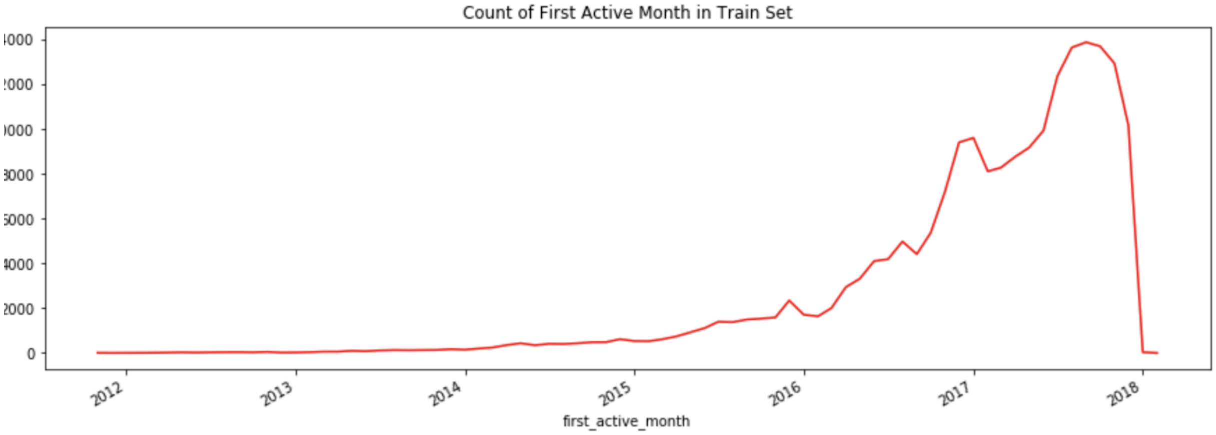
Graphical exploration of above dataset as below:

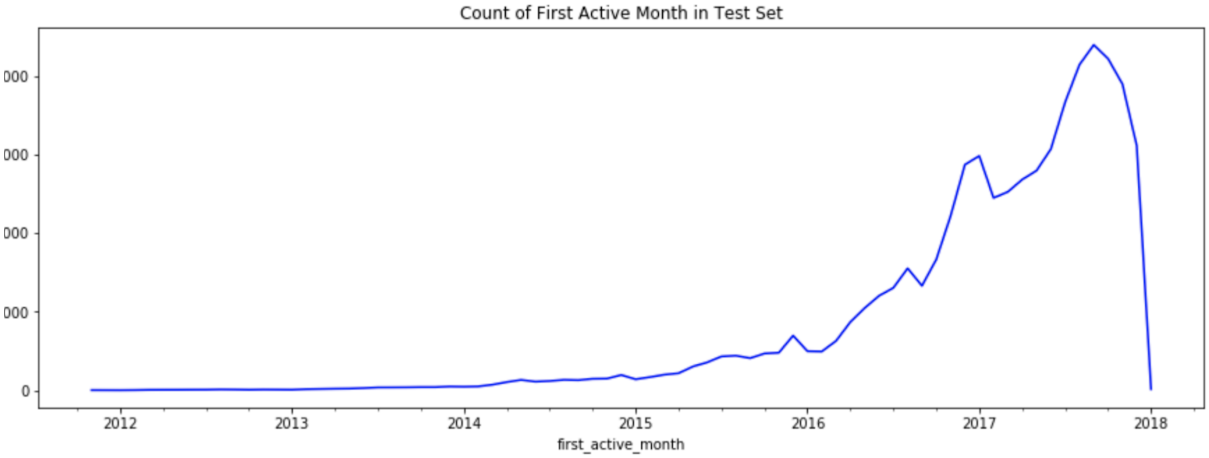




After exploring above it is evident that we need more info to predict loyalty score as the violin plot above appears to be averaging towards zero value.




Count of First Active Month in Train Set
4000
2000
0000
8000
6000
4000
2000
0
2012
2013
2014
2015
2016
2017
2018
first_active_month

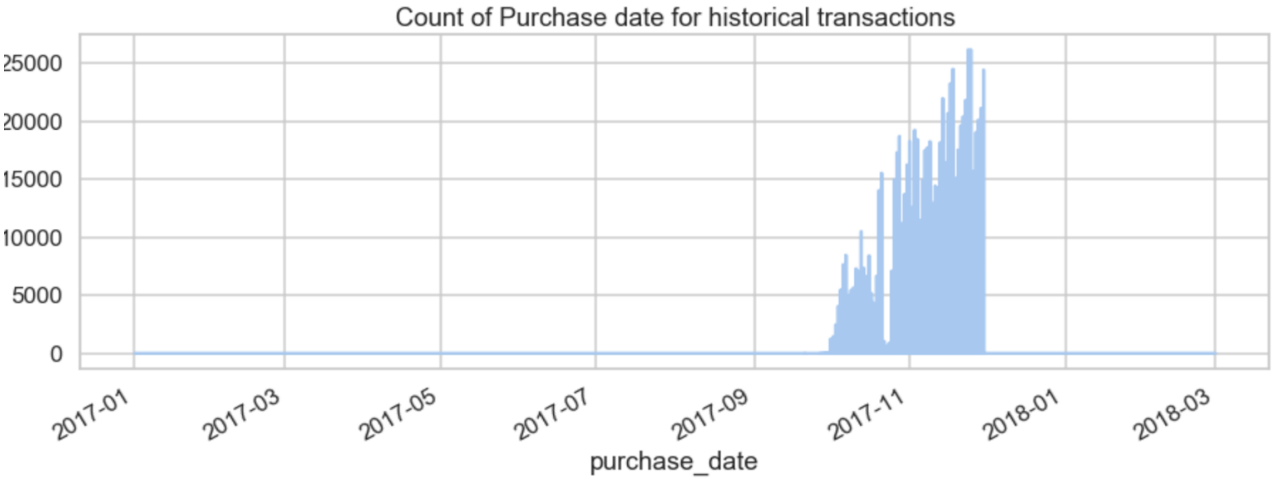

Count of First Active Month in Test Set
000
000
000
000
0
2012
2013
2014
2015
2016
2017
2018
first_active_month

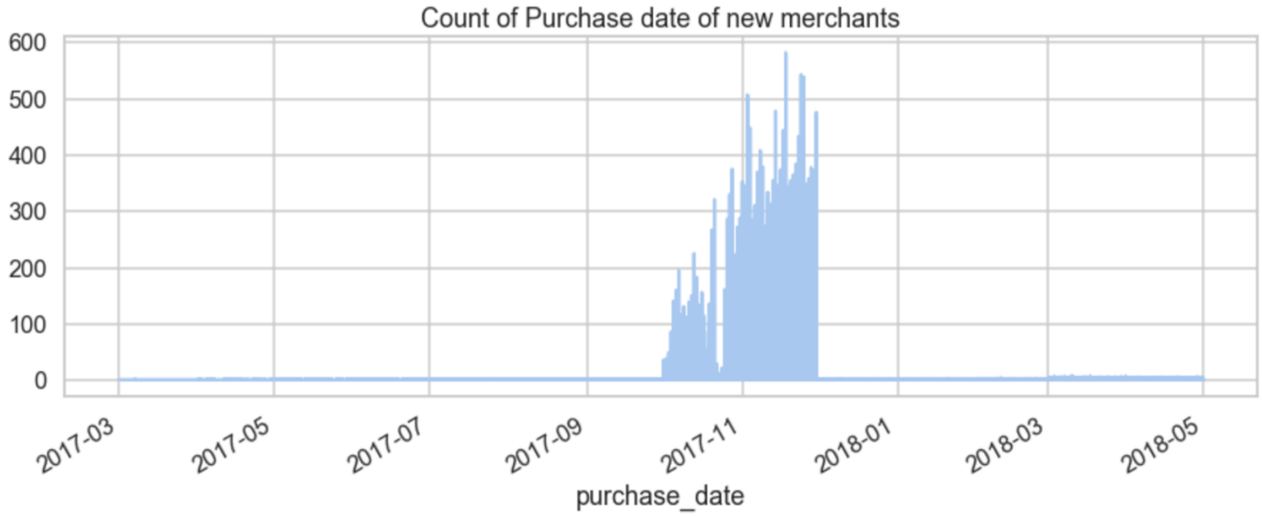
The historical_transactions.csv and new_merchant_transactions.csv files contain information about each card's transactions.historical_transactions.csv contains up to 3 months' worth of transactions for every card at any of the provided merchant_ids.new_merchant_transactions.csv contains the transactions at *new* merchants (merchant_ids that this particular card_id has not yet visited) over a period of two months.

Historical transactions and new merchants contains same information about only with different card ids.

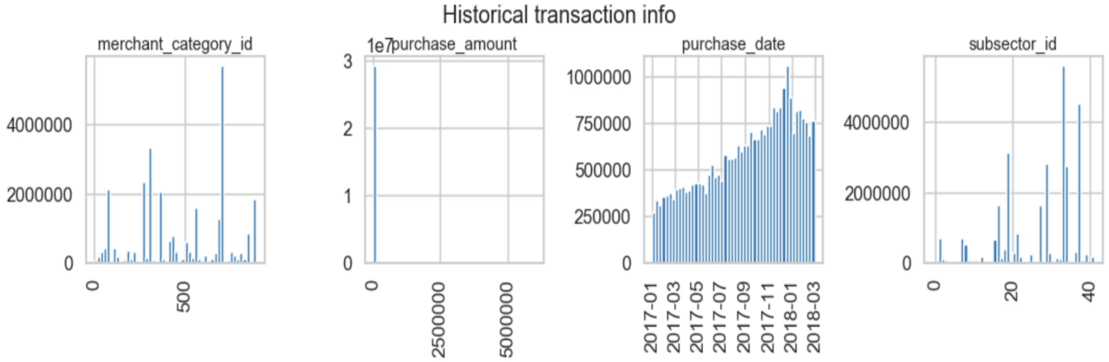
merchants.csv contains aggregate information for each merchant_id represented in the data set.

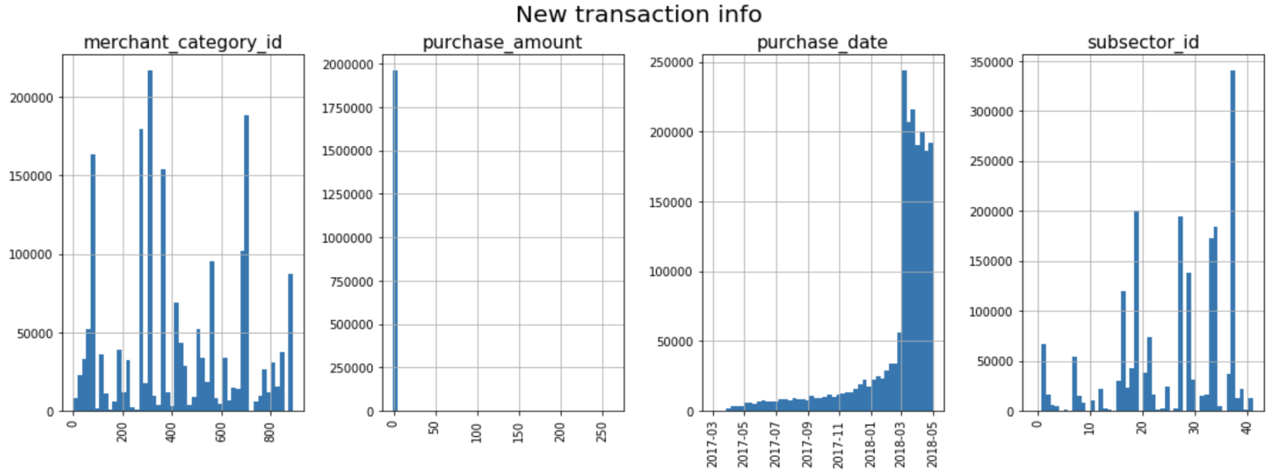
Lets explore the difference between categorical data in new merchants and historical transactions.

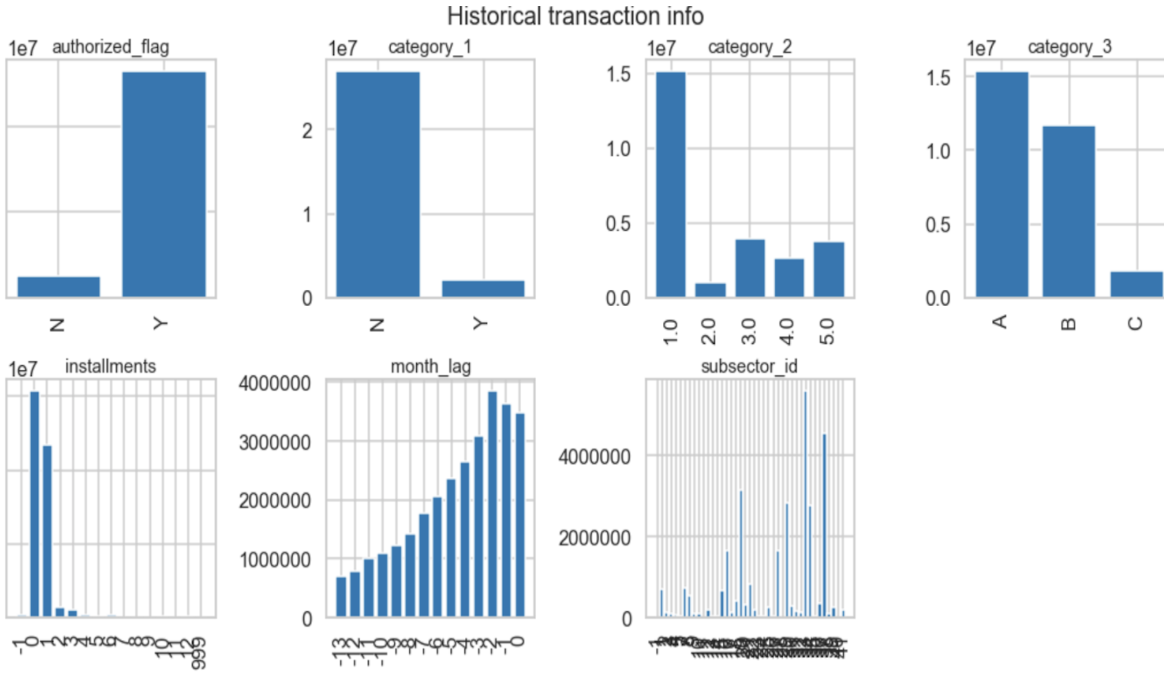


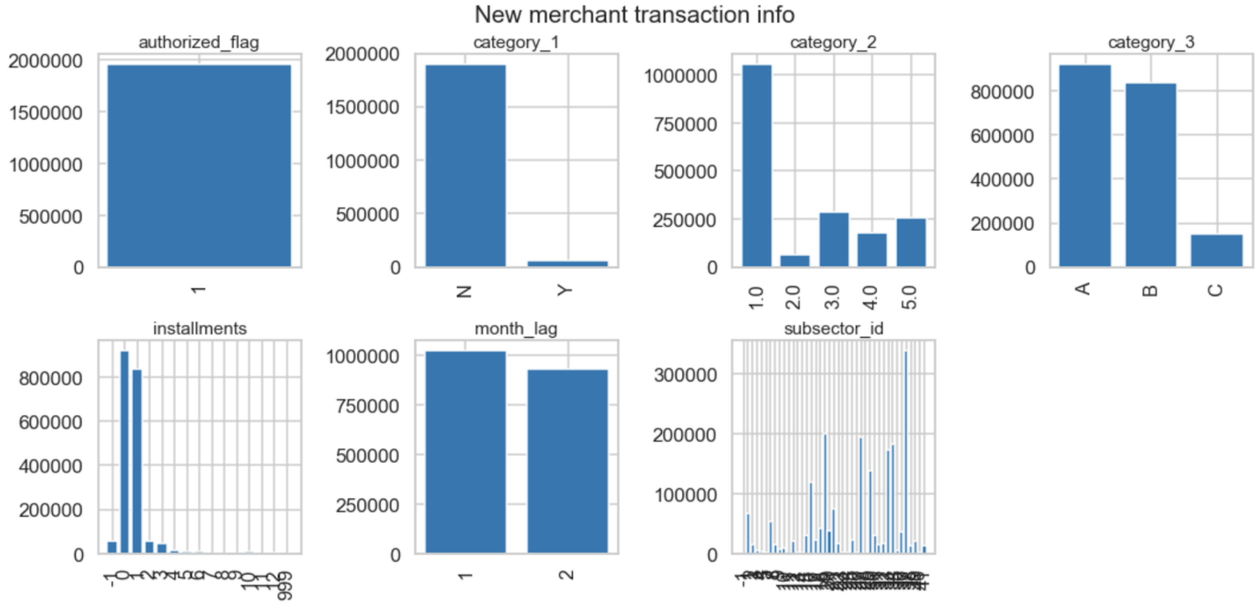

Count of Purchase date of new merchants
600
500
400
300
200
100
0
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
purchase_date

Lets explore visually about the numerical features in both historical and new transaction.






Historical transaction info
1e7 authorized_flag
0
1
2
N
Y
1e7 category_1
0
1
2
N
Y
1e7 category_2
0.0
0.5
1.0
1.5
1.0
2.0
3.0
4.0
5.0
1e7 category_3
0.0
0.5
1.0
1.5
A
B
C
1e7 installments
0.0
0.5
1.0
1.5
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
999
month_lag
0
1000000
2000000
3000000
4000000
-13
-12
-11
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
subsector_id
0
2000000
4000000


New merchant transaction info
authorized_flag
0
500000
1000000
1500000
2000000
1
category_1
0
500000
1000000
1500000
2000000
N
Y
category_2
0
250000
500000
750000
1000000
1.0
2.0
3.0
4.0
5.0
category_3
0
200000
400000
600000
800000
A
B
C
installments
0
200000
400000
600000
800000
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
999
month_lag
0
250000
500000
750000
1000000
1
2
subsector_id
0
100000
200000
300000

Data merging and preprocessing:

As we have analysed various datasets and looking at train.csv and test.csv ,there is need of additional variables to predict the accurate loyalty score.

So we can leverage our objective by combining historical data and new data with the train.csv and test.csv

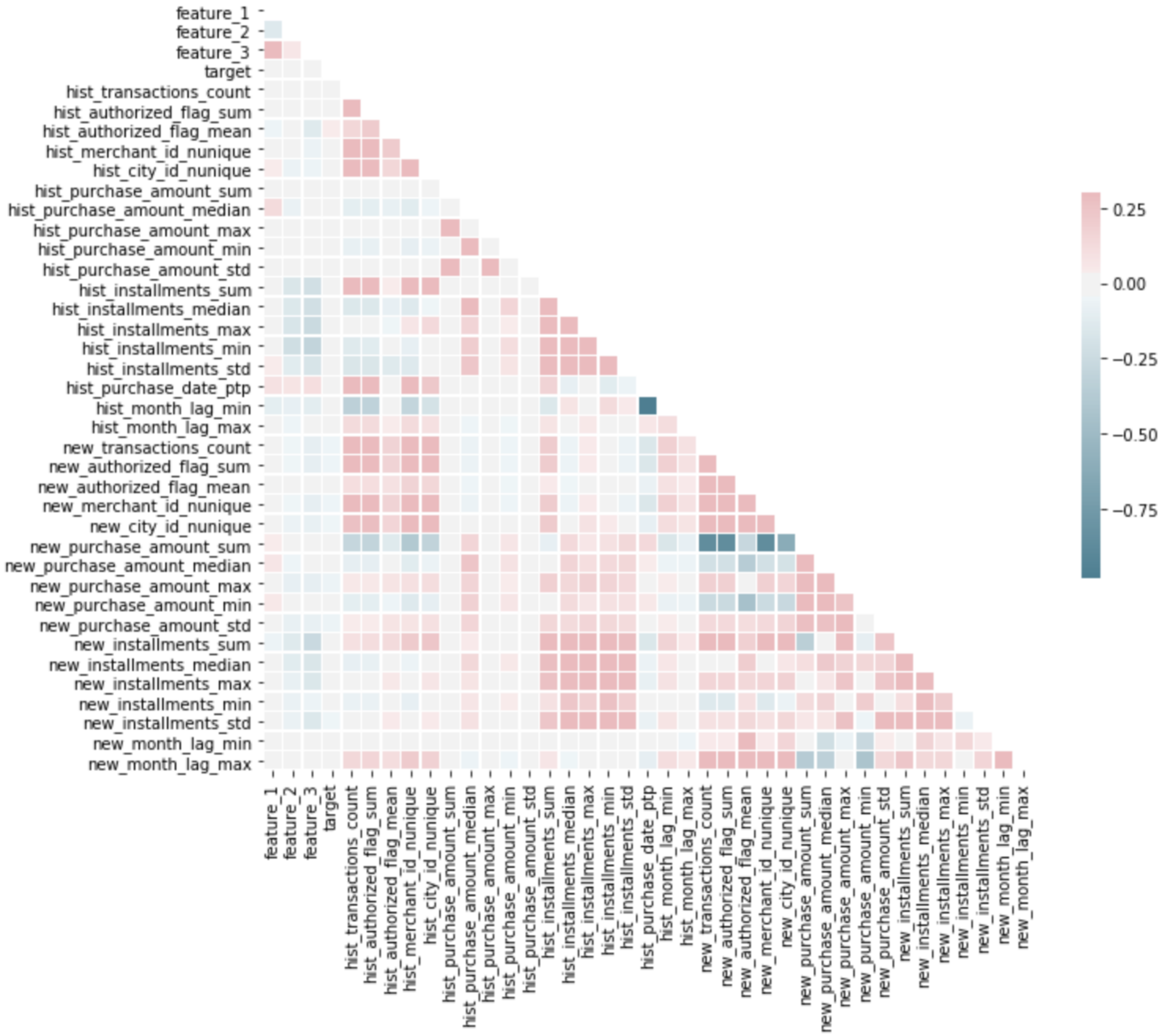
We have aggregated historical transactions as well as new transactions data such as below and then combined with train.csv and test.csv and merged on card_id feature.

```
agg_func = {

        'authorized_flag': ['sum', 'mean'],

        'merchant_id': ['nunique'],

        'city_id': ['nunique'],

        'purchase_amount': ['sum', 'median', 'max', 'min', 'std'],

        'installments': ['sum', 'median', 'max', 'min', 'std'],

        'purchase_date': [np.ptp],

        'month_lag': ['min', 'max']

        }
```

After combining of above.We get training dataframe of size 201917, 41.

After performing co-relations analysis ,we found that there are 13 correlated features so we tried to remove the co-related features from both the train and test datasets.

Co-relation plot is plotted on final train dataframe after combining new transactions with historical transactions as below:



Inferential Statistics - **(Refer jupyter notebook for detailed explanation of test result.)**

Following hypothesis has been tested using python.

H0 hypothesis: There is not a relationship between feature 1,feature 2 and feature 3 with and target

HA hypothesis: There is a relationship between features and target

We can conclude that there is no direct strong correlation between these 3 anonymised features and target indicating we need more data in order to train model to accurately predict target so we have combined various other data such as historical transactions and new transactions data

H0 hypothesis - Mean purchase amount is same between historical and new transactions

HA hypothesis: - Mean purchase amount is different between historical and new transactions

If p <= alpha: Reject null hypothesis that the means are equal.

So we can conclude that the means value of purchase amount is different for historical transactions and new transactions.

H0 hypothesis- Mean installments is same between historical and new transactions

HA hypothesis: - Mean installments is different between historical and new transactions

So we can conclude that the means value of installments is different for historical transations and new transactions.