MerchantCategorisation

December 21, 2021

1 Merchant Categorisation

The objective of this exercise is to be able to segment merchants into significantly different categories based on **key attributes** that we can extract from the available data. The data we have available is that of around 1.5 million transactions across 2 years for 14,351 customers. I make the following general assumption about this data-

1. Each merchant present in this dataset only uses stripe and so we have 100% of their trnsactions in this data.
2. All merchants fall in the same timezone

1.1 Distrubution of merchant features

```
       total_amount_usd    count_txns  count_weekend_txns  \
count      9.149000e+03   9149.000000         9149.000000
mean       2.495361e+04    164.021860           43.104820
std        7.917787e+04    653.750589          188.931226
min        1.589000e+01      6.000000            0.000000
0%         1.589000e+01      6.000000            0.000000
10%        5.193860e+02      7.000000            0.000000
20%        9.792180e+02     10.000000            1.000000
30%        1.632574e+03     14.000000            2.000000
40%        2.667990e+03     19.000000            4.000000
50%        4.281230e+03     28.000000            6.000000
60%        7.232292e+03     42.000000            9.000000
70%        1.289540e+04     73.000000           15.000000
80%        2.400933e+04    127.000000           30.000000
90%        5.528443e+04    297.000000           78.000000
100%       2.369072e+06  25512.000000         7368.000000
max        2.369072e+06  25512.000000         7368.000000

       count_peak_window_txns  avg_txn_amount  transaction_days  \
count             9149.000000     9149.000000       9149.000000
mean               124.903705      310.918030         50.585856
std                501.185820     1124.948013         84.776537
min                  0.000000        2.270000          1.000000
0%                   0.000000        2.270000          1.000000
10%                  5.000000       37.479582          5.800000
```

20%	7.000000	52.376098	7.000000
30%	10.000000	64.748003	9.000000
40%	14.000000	81.192587	13.000000
50%	20.000000	105.272727	18.000000
60%	30.000000	141.166770	26.000000
70%	52.000000	194.441092	40.000000
80%	95.000000	315.467333	67.000000
90%	230.000000	668.826667	134.200000
100%	19030.000000	88874.651667	724.000000
max	19030.000000	88874.651667	724.000000

	activity_duration	days_bw_transactions	transaction_per_day	\
count	9149.000000	9149.000000	9149.000000	
mean	278.924691	12.413963	1.594440	
std	203.849779	17.332963	18.666326	
min	1.000000	0.000596	0.008824	
0%	1.000000	0.000596	0.008824	
10%	24.000000	0.500000	0.033041	
20%	66.000000	1.215197	0.054422	
30%	124.000000	2.128536	0.082853	
40%	186.000000	3.484444	0.123457	
50%	255.000000	5.515152	0.188742	
60%	328.000000	8.521739	0.299922	
70%	401.000000	12.926374	0.489230	
80%	481.000000	19.862112	0.853008	
90%	578.000000	33.337255	2.049151	
100%	730.000000	136.000000	1679.000000	
max	730.000000	136.000000	1679.000000	

	perc_weekend_txns	perc_peak_window_txns
count	9149.000000	9149.000000
mean	0.240665	0.748583
std	0.183396	0.202662
min	0.000000	0.000000
0%	0.000000	0.000000
10%	0.000000	0.500000
20%	0.088235	0.631579
30%	0.142857	0.700000
40%	0.181818	0.745859
50%	0.222826	0.782609
60%	0.264099	0.823529
70%	0.301622	0.862069
80%	0.354327	0.909091
90%	0.452814	0.988036
100%	1.000000	1.000000
max	1.000000	1.000000

The above table helps us get a sense of the distribution of several key merchant features. Merchants with less than 5 transactions have been skipped.The following features have been computed at a merchant level -

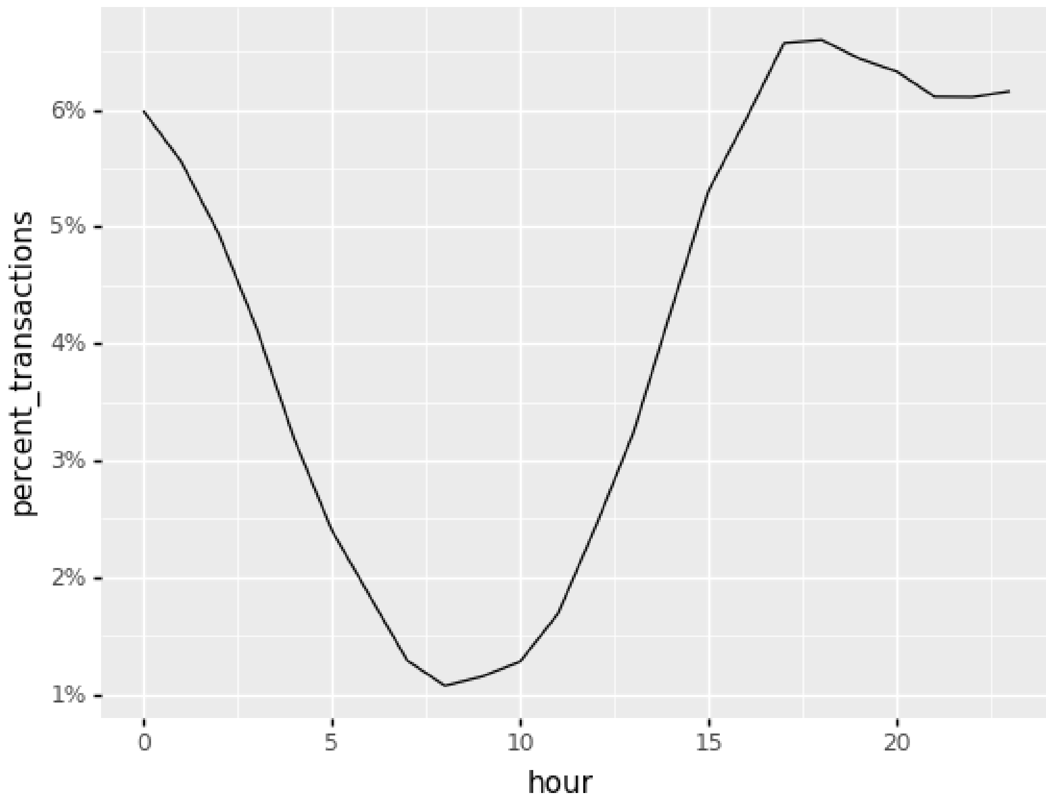
1. count_txns : Total number of transactions
2. count_weekend_txns : Total number of transactions that happened on a weekend
3. count_peak_window_txns : Count of transactions that happened during the daily window of 3pm to 3am (see below pn how i came up with this)
4. avg_txn_amount : Average transaction amount
5. transaction_days : number of unique days on which the merchant transacted
6. activity_duration : time period in days from first to last transaction for the merchant
7. days_bw_transactions : average days between consecutive transactions for the merchant
8. transaction_per_day : Average transactions per day within the activity_duration
9. perc_weekend_txns : Percentage of transactions that happened on the weekend
10. perc_peak_window_txns : Percentage of transactions that happened during the daily window of 3pm to 3am

I did not venture into the time distribution of trnsactions beyond weekly since we have limited data. If we had data for a few more years, we could make out some quarterly/monthly trends as well. I have also ignored merchants with less than 5 transactions within the 2 year period since getting statistically relevant features for these merchants would not be possible and we should probably wait for more data on them before we categorize them.

A few trends jump out by looking at the distributions.

1. The average ticket size is log normally distributed i.e. around 80% of the population has an average of ~300 usd, but there is a fat tail wehre a few merchants have really high average transation sizes.
2. The average days between transactions has a similar disrtibution where the median average days bw consecutive transactions is around 5 days.
3. Almost half the merchants have 80% or more of their transactions happening in the 12 hour window from 3pm to 3am.

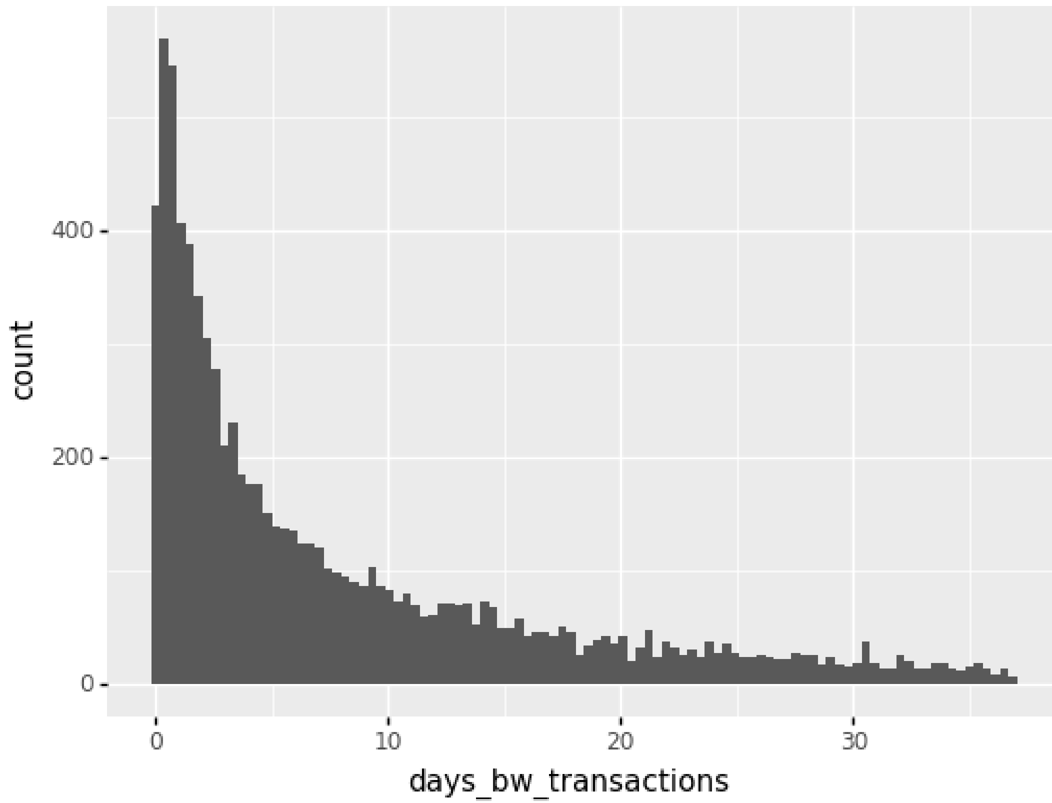
1.1.1 Distribution of transactions across the day




```
<ggplot: (-9223371888873758072)>
```

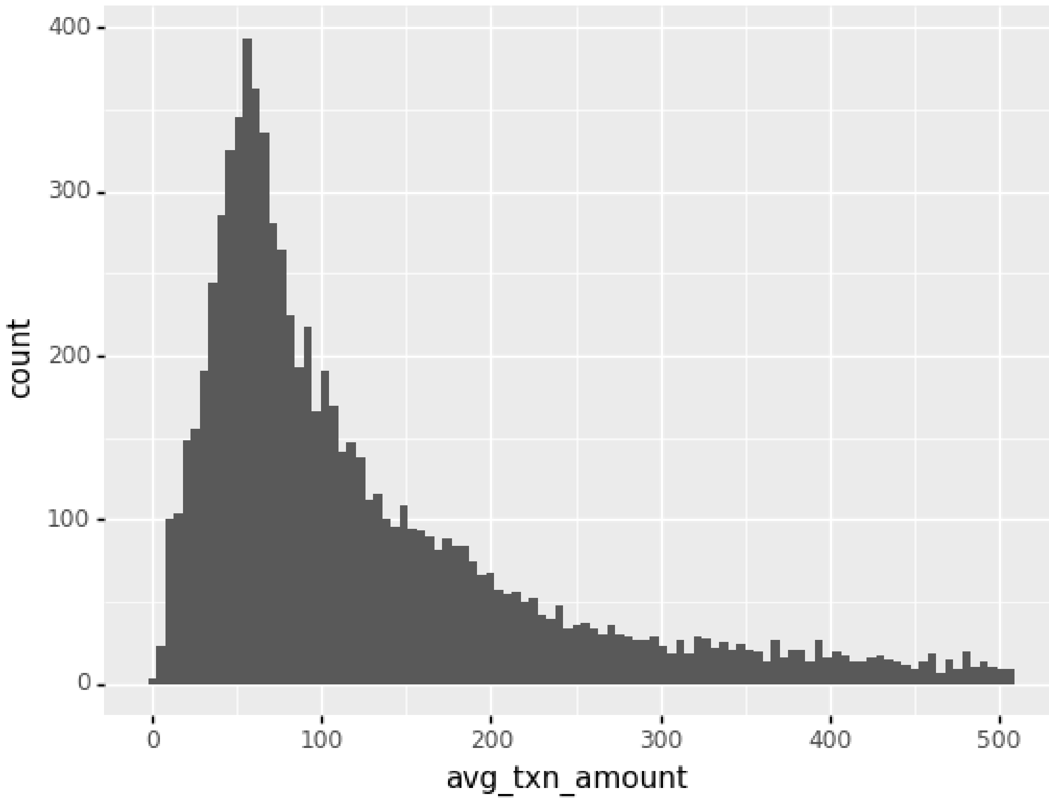
There is significantly more activity in the 12 hour window from 1500 to 0300 than the rest of the day. It would be interesting to see if this is contributed by a specific category of merchants.

1.1.2 Distribution of average days between consecutive transactions




```
<ggplot: (-9223371888872171300)>
```

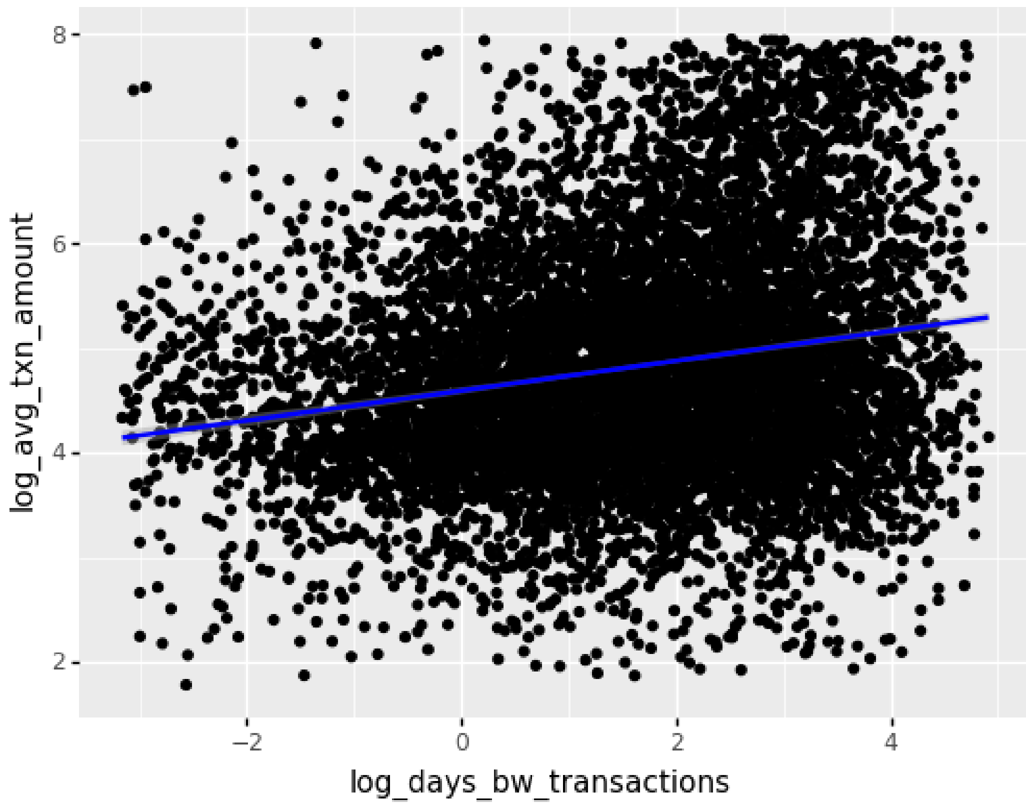
1.1.3 Distribution of average transaction size in dollars




```
<ggplot: (-9223371888869421300)>
```

2 Relation between average ticket size and frequency

```
(8912, 2)
```

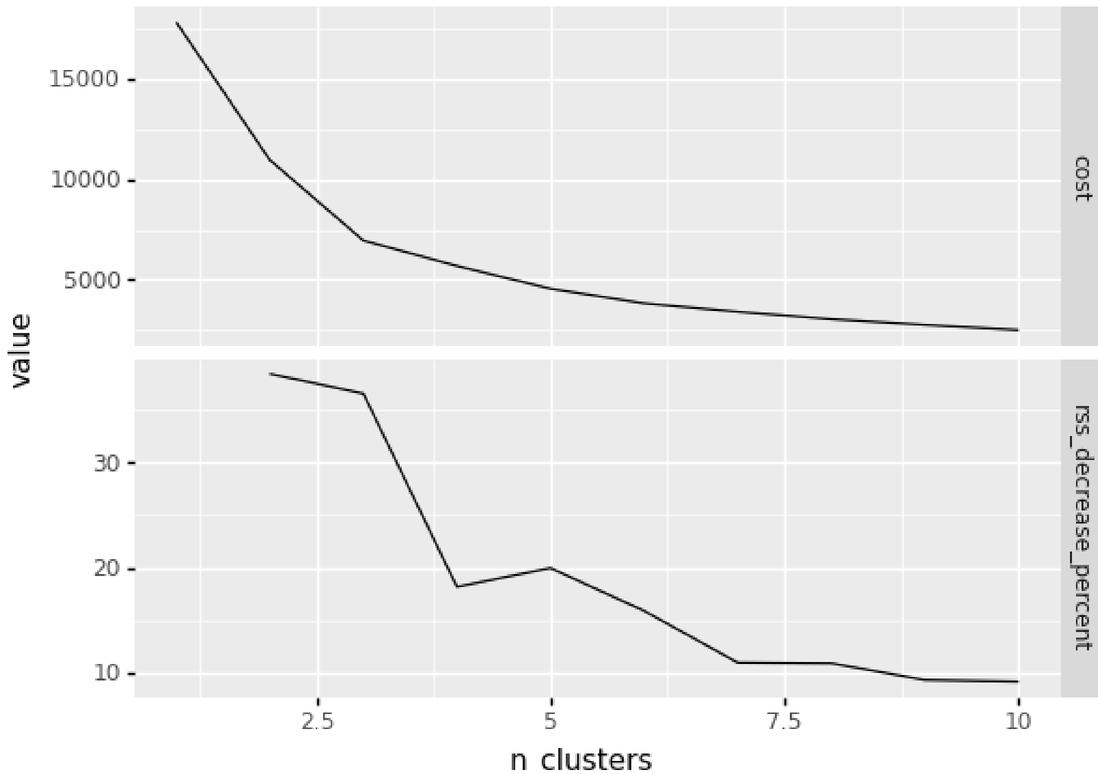



```
<ggplot: (-9223371888868388860)>
```

Above is a scatter plot between the average ticket size and the average fays between transactions for all merchants. Both axis are on a log scale. CLearly There is a significant correlation of the frequency(inverse of average days bw transactions) of transactions with ticket size. As frequency increases, merchants are likely to be lower ATS. The relationship seems to be almost linear albeit with some heteroskedasticity as the variation of ticket size is much higher amoug merchants with lower transaction frequencies(higher average days between transactions).

3 Running a simple clustering model

From the above exploratory data analysis, the features that make the most business sense would be the average transactions size of the customer, captured by the average dollar transaction size , and the frequency of trnsactions, captured by the average days between transactions.

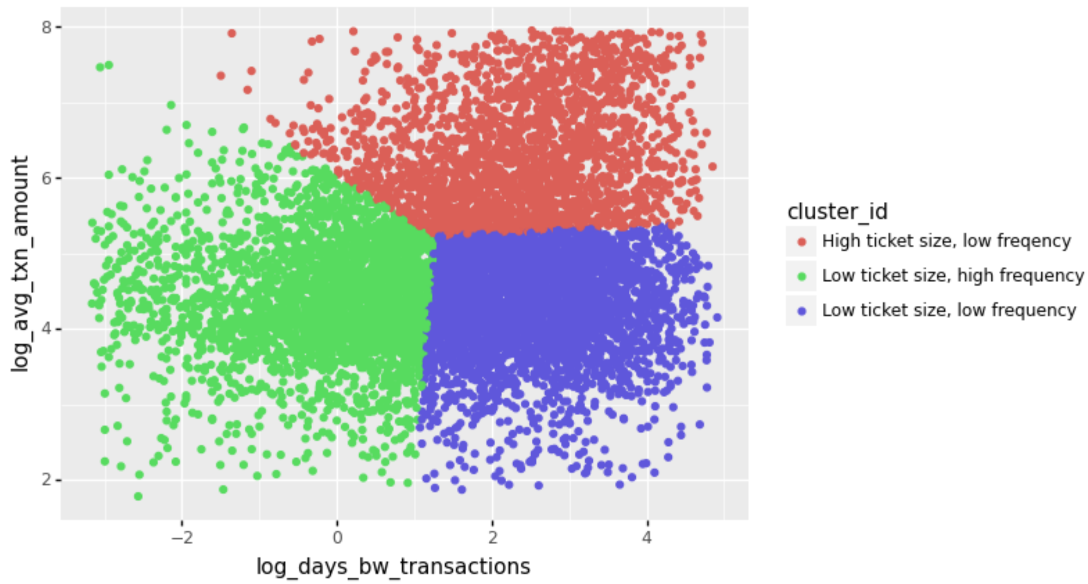



```
<ggplot: (-9223371888873757956)>
```

3.0.1 Choosing the optimum number of clusters

Beyond 3 clusters, the decrease in the unexplained variance(in pecentage terms) drops sharply. So I have chosen 3 clusters to be the optimal for this case.

```
KMeans(max_iter=500, n_clusters=3, random_state=42)
```

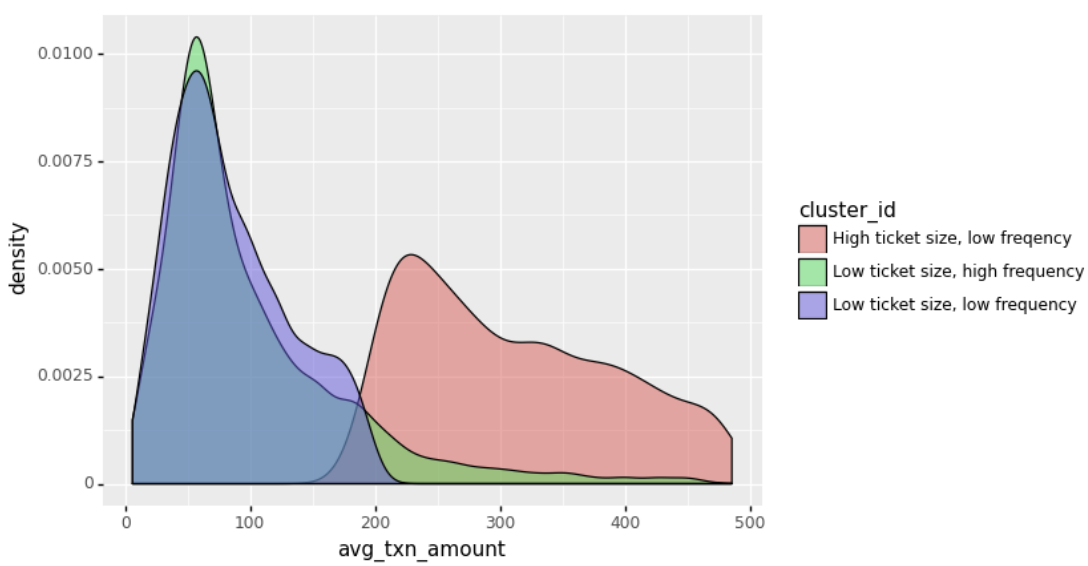


`<ggplot: (-9223371888873534716)>`

Visualising the clusters in log space, the clusters represent 3 distinct categories -

1. Merchants who transact frequently and have low average days between transactions.
2. Merchants who transact infrequently but have a higher average transaction size when they transact.
3. Merchants who transact infrequently and also have a lower average transaction size when they transact.

The below density charts highlight the differences in these segments clearly -



<ggplot: (-9223371888873546340)>

<ggplot: (-9223371888873523908)>

3.1 Limitations

- Have not explored the distributions of payment amounts and the frequency of payments, just used averages, further information can be derived from nuances in the distributions and joint distributions. Eg subscriptions etc.
- Have not explored the dimension of how merchants are different in the time window in which hey operate. Almost half the merchants have 80% of more trnsactions in the time window between 3pm to 3am. This can be a very valuable dimension for us.

3.2 Possible use cases

The different segments need a different kind of service from stripe. For example, merchants where the frequency of transactions are high but the ticket size is low, we could work on automating and reducing any operational costs since itheyt would be incurred for each transaction. We can also be more confortable with the risk in these cases since the ticket sizes are low and any single transaction being fraudulent does not have a very high cost realtive to the revenue generated by the merchant for us. Similarly merchants with high average ticket sizes and do not transactneed stricter fraud checks.