Analysis of multicategory online shopping mall and its users

Category

- Project overview and organization
- Procedures and methods of project implementation
  - Hypothesis about the problem
  - Hypothesis test and problem definition
  - Behavior analysis of customers
- Consequence of the project
- Self-assessment

Project overview and organization

Exploration of data and set hypothesis	Hypothesis test and problem definition	Behavior analysis of customers
05-Jan ~ 06-Jan	07-Jan ~ 08-Jan	09-Jan ~ 12-Jan
Sampling EDA	Feature Engineering Pearson Correlation Analysis	Clustering Pearson Correlation Analysis Visualization
SQL (SSMS), Python, Visual Studio, Colab	Python, Colab	Python, Tableau, Colab
Does this mall fail to retain users?	▪ Failure of retention ▪ Failure to sell multi-category	▪ Focus on high-involvement products ▪ Suggestion of relevant marketing strategy

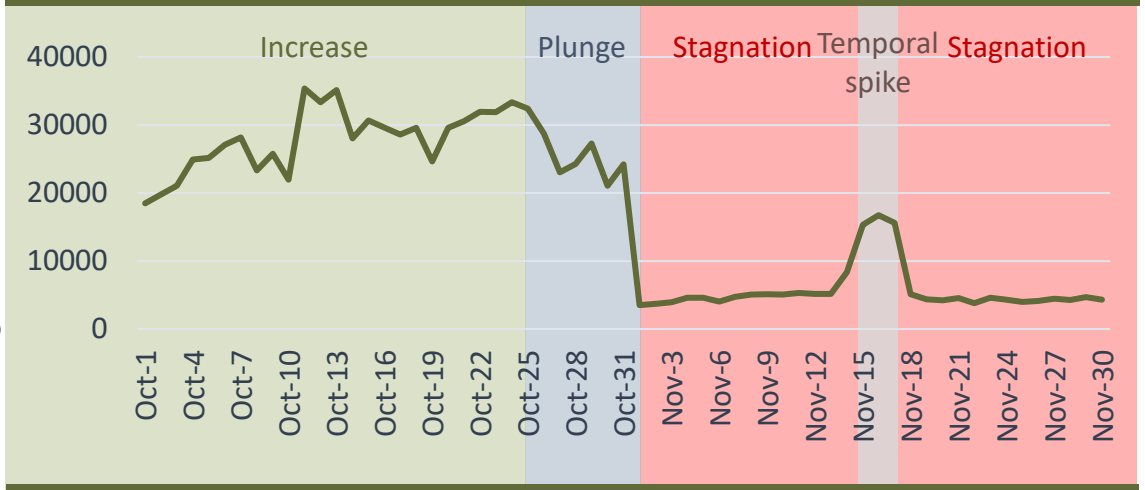
Hypothesis about the problem

Failure to retention of customers is assumed

Analysis on 9,780 users

	2019-Oct	# events		2019-Nov	# events	
Total	100%	42,448,764		100%	67,501,979	
view	100%	40,750,813	96%	100%	63,451,860	94%
cart	100%	848,975	2%	100%	2,700,079	4%
Other	100%	848,975	2%	100%	675,020	1%
Sample	2.0%	848,980		0.3%	172810	
view	2.0%	820,656	97%	0.3%	162,140	94%
cart	1.8%	15,093	2%	0.3%	8,210	5%
purchase	1.6%	13,231	2%	0.4%	2,460	1%
Non-Sample	98.0%	41,599,784		99.7%	67,329,169	
view	98.0%	39,930,157	96%	99.7%	63,289,720	94%
cart	98.2%	833,882	2%	99.7%	2,691,869	4%
Other	98.4%	835,744	2%	99.6%	672,560	1%

Trend of number of events from Oct. to Nov.



- Declined in # of events of users in Oct
  - With high fluctuation, the number of events increased until 25th Oct
  - The number of events plunged since 25th Oct
- Settled in about 5000 in Nov
  - Temporal spike from 15th to 17th in Nov
  - This spike seems to be due to promotion and fail to retain customers

Since it was impossible to load 5GB and 9GB data with the memory provided by Colab, a sample was extracted and analyzed. I wanted to extract 10,000 users, but when I set the user base in SSMS and extracted it with the TABLESAMPLE (ROWS = 10000) command, 9,780 users came out. November was also extracted according to the same number of people as in October.

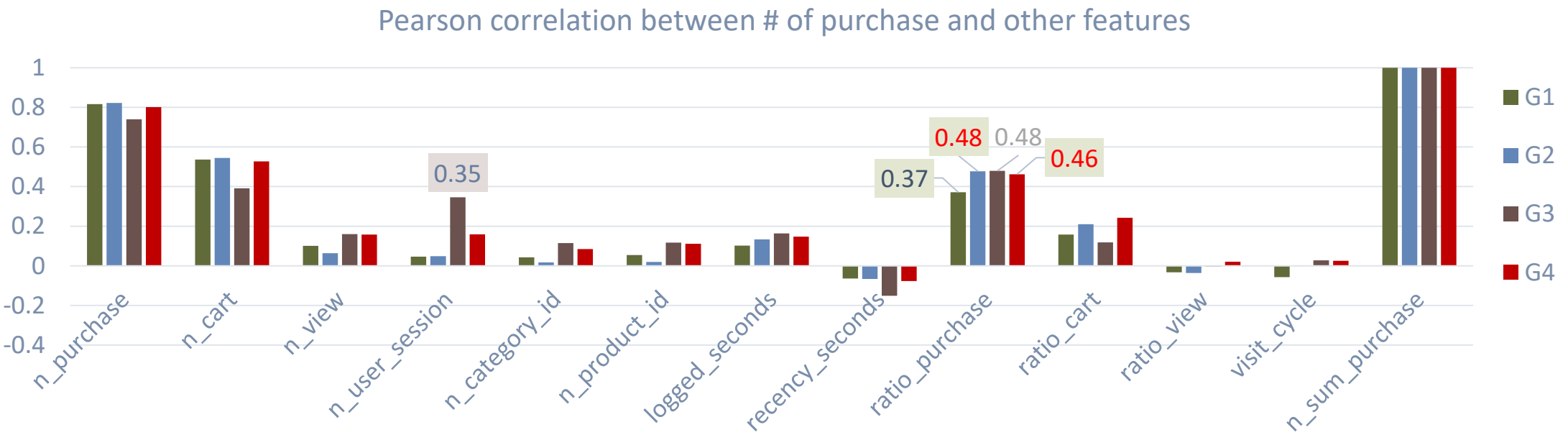
In order to analyze the characteristics of customers who maintain use, they are processed into characteristics aggregated by user

Segmentation in four types of time period

Gr-oup	Period in 2019	Trend in the size of event
G1	10/1 ~ 10/25	Increase
G2	10/25 ~ 10/30	Plunge
G3	11/15 ~ 11/17	Temporal Spike
G4	11/1 ~ 11/15 11/18~11/31	Stagnation

Feature	Description
n_purchase	The number of purchase of a user
n_cart	The number of cart of a user
n_view	The number of view of a user
n_user_session	The number of log-in of a user
n_category_id	The number of categories a user viewed or put in cart or purchased
n_product_id	The number of products a user viewed or put in cart or purchased
logged_seconds	Difference between the first logged-in time and the last logged-in time (seconds)
recency_seconds	Difference between the last time of the analysis and the last time the user logged in (seconds)
ratio_purchase	Ratio of the purchase event (n_purchase / n_user_session)
ratio_cart	Ratio of the cart event (n_cart/ n_user_session)
ratio_view	Ratio of view events (n_view / n_user_session)
visit_cycle	Visit cycle (logged_seconds / n_user_session)
n_sum_purchase	Total value of a user's purchase

Users who used mall in stagnation (G4) are target persona to analyze

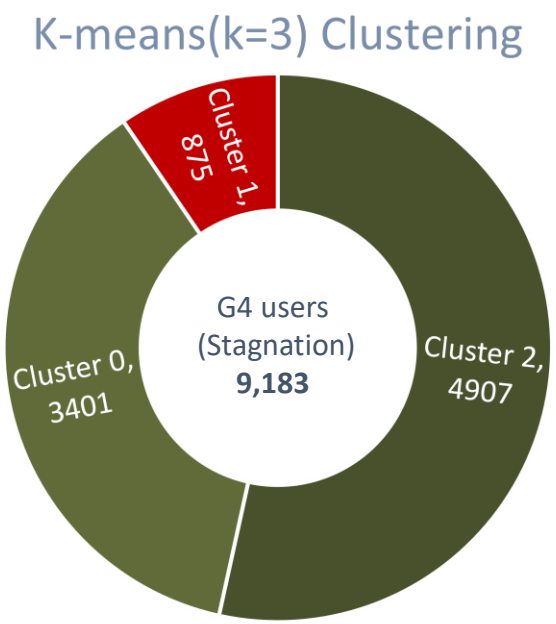


- Since the number of the events decreased in Oct, the correlation of "purchase ratio" and "sum of the purchase" increased. This means users who had intends to purchase increased more than those who just look around and don't purchase
- In the mid of November, the high correlation of the sum of purchase and the number of user session is presumed to be due to promotion

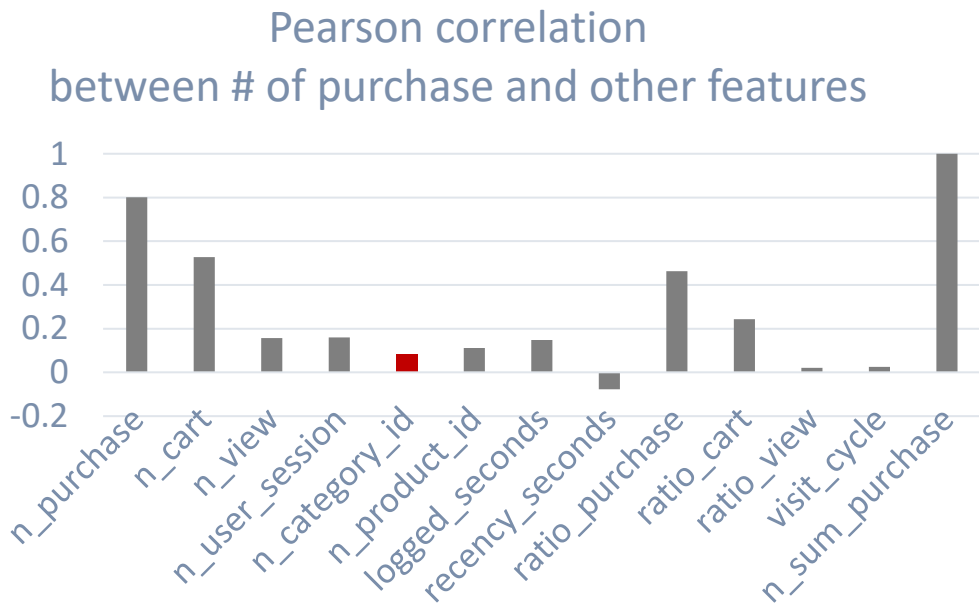
The users of stagnation in November are plausible to continue to use this mall onward because of the high correlation between the 'ratio of purchase' and the 'sum of purchase value'

Behavior analysis of customers

Little correlation between the diversity of categories and purchases, …it is necessary to find the main categories that customers purchase

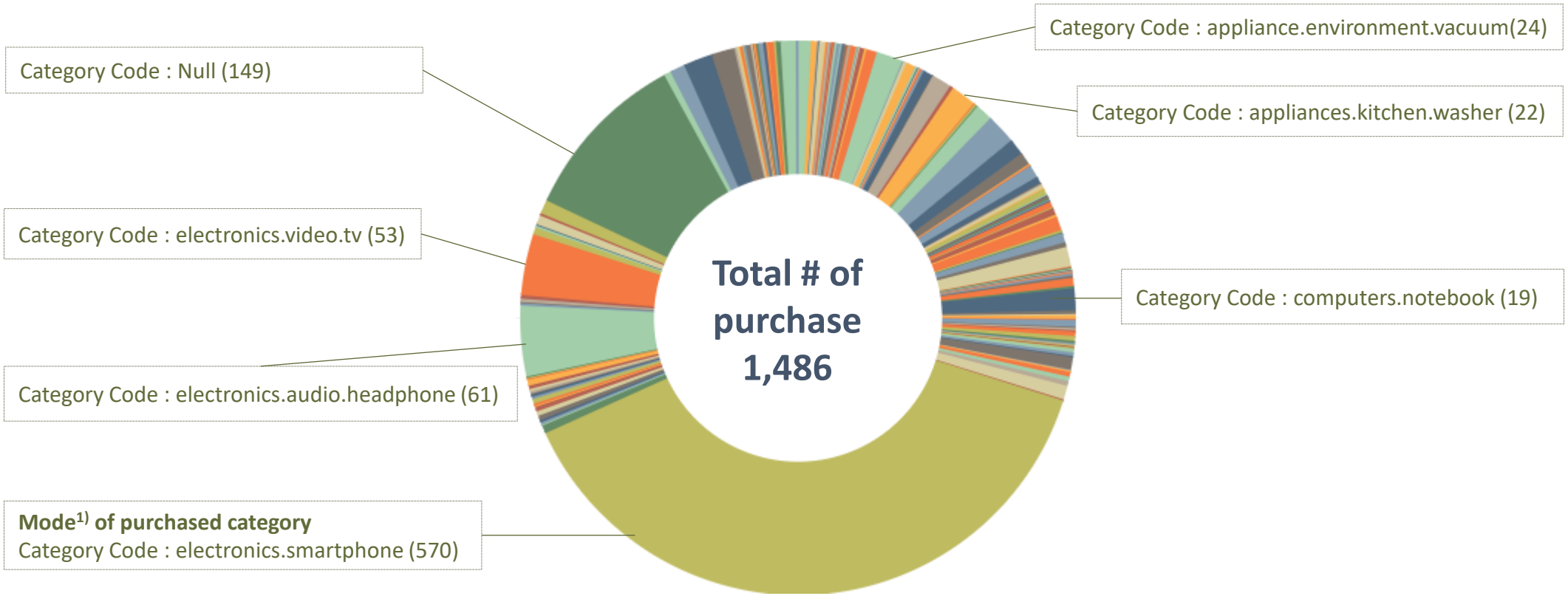


Clusters 0 and 2 have no purchase history, so the analysis target is reduced to Cluster 1



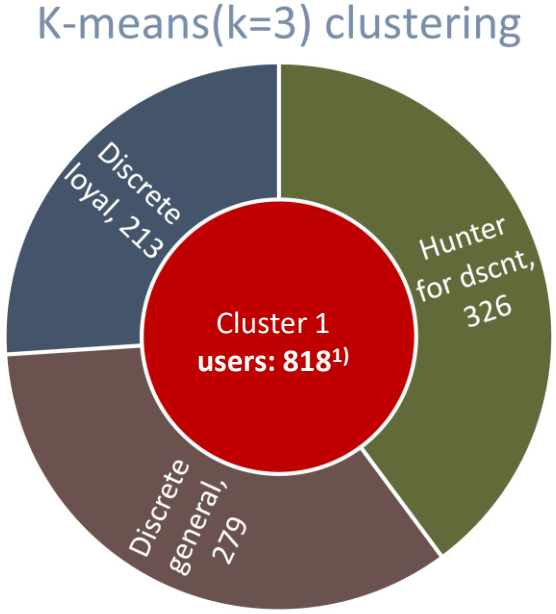
Maintaining multi-category inventory may be disadvantage because viewing multiple categories does not increase the purchased value

Consumers primarily buy high-involvement products



Mode of purchased category means most frequently purchased category

Users are segmented into three parts, of which discrete loyal group spends 1 hour and 10 minutes per session



	Discrete loyal customer	Hunter for discount	Discrete general customer
Recency	5 days	2 weeks	1 week
Frequency	Purchased 3 times	Purchased 1 time	Purchased 1 time
Monetary[2)]	857.04	372.71	125.14
Involvement	**1 hour 10 minutes**	9 minutes	1 hour
	x 213	x 326	x 279
Contribution to revenue[2)3)]	54% (182,550)	36% (121,503)	10% (34,914)

1) Of the 875 users in Cluster 0, 56 users who did not purchase at all were excluded
2) The units of Monetary and Contribution to revenue are not indicated because they are not precise. (However, the same units are used)
3) Sales refers to the percentage of customer groups in total sales, not per customer

Consequence of the project

It is necessary to understand the customers who purchase high-involvement products and to devise a marketing strategy

- Although it is a shopping mall that sells multi-categories, the categories that customers actually purchase are high-involvement products such as electronic products and electronic furniture
- Need to cluster with RFM + I (involvement), and develop marketing strategy according to customer tendency
  - For customers with a long navigation process → cautious customers: need to provide a lot of information
  - For customers with a short navigation process → customers with quick decisions: only key information needs to be prominently provided
- Since users have already recognized that it is a good platform for purchasing electronic products, a strategy to further specialize the UI by providing convenient functions for purchasing electronic products is effective
- When competitors who sell multicategories, such as Amazon, already have an edge in the market, focusing on a specific category is a way to attract new users
- Since high-involvement products do not have a high frequency of purchase, it is important to not only increase the number of users, but also to continuously update the shopping mall so that it can be used for a long time to secure LTV

Self-assessment

Regrets, learnings and achievements

- In the case of multi-category sales, it would be possible to determine at the EDA level whether a particular category was selling the most.
- If I had used Tablo from the beginning, the EDA process would have been faster.
- I don't use statistics much, so I concluded what anyone with a lot of domain knowledge would have already known.
- If we had targeted users with at least one purchase history from the beginning, the clustering steps would have been reduced
- The increase in data size from October (5GB) to November (9GB) can be expected to indicate an increase in users, but it would be a mistake to judge that the number of events decreased in November. It would have been nice to include not only the number of users, but also a certain percentage of the total data collected and compare them
- When visualizing in PowerPoint, using Excel and PowerPoint charts is easier and faster than Matthew and Seaborn.
- Succeeded in sampling
- Persona setting logic is appropriate
- I discovered the possibility of wasting inventory costs and suggested the strategic direction of the shopping mall.