# Analyzing Credit Card Usage Patterns with Deep Clustering

Keonwoo Kim
*School of Industrial Engineering*
*UNIST*
keonwookim@unist.ac.kr

Youngseok Song
*School of Industrial Engineering*
*UNIST*
mrsys@unist.ac.kr

Youngbin Lee
*School of Industrial Engineering*
*UNIST*
young@unist.ac.kr

***Abstract***

**To identify the card consumption patterns, BC Card's consumer data is grouped by region, gender, income from January 2019 to April 2020. In this project, using quantitatively measured household data, we divide consumer groups by characteristics and identify changes and patterns among them. It is also intended to obtain meaningful information about the card industry through consumption forecasting. The data received from the BC card goes through a data cleansing process first. After removing missing values and outliers in card consumption data organized by month, new categories were created and reclassified. After that, clustering was performed using the N2D method, and the number of clusters considered to be the most representative of the characteristics was determined. As a last step, we tried to obtain meaningful information about clusters using the pie chart, transition matrix and glide paths. Based on the characteristics and graphs obtained in this way, conclusions were made and limitations were also described.**

## I. INTRODUCTION

The purpose of this study is to analyze the various degrees of consumption by analyzing the card consumption history. In addition, the consumption shock rippled by social risks and crises caused by COVID-19 is analyzed by age and gender. It goes through the process of classifying card consumption patterns by grouping BC card consumer data from January 2019 to April 2020 according to region, gender, and income. By classifying consumption patterns, we want to quantitatively measure household types by observing changes between patterns and monthly changes in consumption. In addition, it is to analyze consumption before and after the spread of the corona virus and to help predict consumption. First, missing values and outliers are removed from the data. After data cleansing, industries are reclassified according to criteria arbitrarily set by the experimenter and grouped by region and customer information to see consumption rates by industry. After analyzing consumption patterns through clustering, characteristics of each cluster, transition matrix, and monthly cluster movements were analyzed.

This study was conducted using the credit/debit card consumption data of BC Card's customers. Consumption data used by approximately 17 million customers from January 2019 to April 2020. It was provided by processing customer card consumption data and transaction data for each card approval, which are source data generated when consumers make payments at card merchants. The rows are approximately 90M, and the columns have 14 features. Table 1 summarizes the detailed elements of the data. Looking at the table from the top, year and month, with data used between January 2019 and April 2020. In customer type, If the card usage area and the customer address are the same, they are classified as residents, otherwise they are classified as tourists, and they are displayed as local-resident and local-tourist. The locations of affiliated stores registered in BC Card are classified by metropolitan city, city, gun-gu, and administrative-dong units. And in Business classification the three categories of industries are listed according to BC Card's business classification standards.

The customer attributes that need to be explained in particular are the Household life cycle and Income estimate. First, household life cycle is an estimate based on the customer's basic profile (age, occupation, income, etc.), card characteristics (check/credit, amount used, etc.), preferred service, and lifestyle (classified according to consumption share by industry). They are grouped into 5 categories as shown in the Table 1. Second is income estimate. In order to

evaluate an individual's ability to make payments, they are classified according to the nature of the customer ledger (whether or not they have a credit or debit card) and occupation. And also, BC Card usage amount and external information (type of residence, land price per square feet, sale price/multi-card performance, loan, delinquency information, etc.) were used to estimate income. They were classified as B1 to B11, starting with 1 for the lowest income level.

**TABLE 1**

| Item | Details | Data Value |
|---|---|---|
| Year and Month | | 2019.01 to 2020.04 |
| Customer type | | local-resident/local-tourist |
| Card use location | locations of affiliated store 1 | metropolitan city |
| | locations of affiliated store 2 | city , gun-gu |
| | locations of affiliated store 3 | administrative-dong |
| Business classification | category names of business large | |
| | category names of business medium | |
| | category names of business small | |
| Customer Attribute | Gender | Male, Female |
| | Age | under 20, 20s, 30s, 40s, 50s, 60s or older. |
| | Household life cycle | single-person households, elderly households, households with adult children, households with newlywed infants, and households with elementary, middle and high school children. |
| | Income estimate | B1 ~ B11 |
| Transactions | Amount of usage | |
| | Number of transactions | |

## I I. Literature Review

### *A. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding*

$$C = Fc(Fm(Fa(X))) \quad (1)$$

Deep learning techniques has widely been used on clustering methods such as autoencoder or neural networks. However, in this paper, N2D applies on shallow clustering algorithms unlike usual ones. N2D model consists of three clustering methods which are auto encoder, manifold learner and simple clustering algorithm. (1) shows N2D algorithm's main principle which shows clustering procedures given input data 'X'. First of all, by using auto encoder method, data is embedded into latent vector. After embedding the data, the dimension is reduced by using manifold learner, which treats specific algorithms 'Uniform Manifold Approximation and Project' (UMAP). Finally, universal clustering technique is used, 'Gaussian Mixture Method' (GMM) in this paper.

As mentioned above, this paper mainly focuses on two different manifold learning methods which are autoencoder and UMAP. The first one is a deep neural network which consists of two parts, encoder and decoder. Encoder compresses the input data into a lower dimensional space which means that it maps the input data 'X' into a new smaller feature vector, so called 'latent-vector'. Decoder works the opposite method of the encoder. It maps the learned feature vector into original vector, which refers to make 'latent-vector' to have same dimension with original input data's one. The limitation on this model is that it does not preserve the distances of data in the representation that they learn. UMAP was selected as a second manifold learner among several manifold learners such as Isomap, t-SNE and UMAP, since it shows best performances. Unlike UMAP, t-SNE has difficult in large size data and does not preserve global structure. Also, similar to Isomap, UMAP uses a k-neighbour based graph algorithm but it can construct a weighted k-neighbour graph at higher level. One of the important characteristics on UMAP is modeling the manifold with a fuzzy topological structure which refers that everything is a matter of degree and does not have deterministic value but only analog with the infinite concentration of gray between black and white.

The final clustering algorithm is GMM, where each component has its own general covariance matrix, and it has 'c' components which refer the number of clusters. Its idea of clustering is to represent the probability distribution on real world into mixture of 'k' number of Gaussian Distributions.

Accuracy (ACC) and Normalized Mutual Information (NMI) are used as evaluation metrics on clustering results. Autoencoder's hyperparameter was set as 1000 for epochs, Adam for optimizer and ReLU for activation function.

The results on experiments which compare performance on diverse tests MNIST, USPS, Fashion, pendigits and HAR are that N2D with UMAP for second manifold learner shows SOTA on the most parts based on ACC and NMI. Learning the local manifold of an autoencoded embedding, while also keeping global structure like UMAP, can be more appropriate on several problems to find out the well classified clusters.

## III. Method

### *A. Remove outliers and missing values*

In the data cleansing process, we first checked and removed the row filled with 'x'. After checking the row with the 'x' value in the annual average income estimate, the row was deleted. Considering a case that exists in other categories (like age, household life cycle) as well, we examined the value of 'x' and deleted the row with the value of 'x'. Then, the rows with missing values were removed using python function. As

a result, it was found that about 2% of the total rows were removed by the above process.

Outlier values were calculated based on the category names of small business. Since we thought that the unit amount for each sub-category reflects the characteristics of the industry, we set it as a criterion for checking outliers. Dividing the total amount used by the number of use cases, we looked at the distribution of the unit amount by the category names of small business . It is set to remove less than 5% of the total data while calculating and excluding values that are very out of range.

### *B. Category mapping*

**TABLE 2**

| New Category Names | |
|---|---|
| Restaurant | Hospital |
| Transportation | Grocery |
| Entertainment | Service |
| Clothes | Vehicle |
| Insurance | Education |
| Travel | Furniture and Appliance |
| Mart | Retail |
| E-Commerce | Etc. |

In the original column of BC Card customer data, there are ‘Small, Middle and Large divisions’ which grouped the data for specific expenditures. Using these divisions to analyze, it takes a lot of cost and time. Thus, we created 16 number of new classifications of expenditures arbitrarily. TABEL 2 shows the results of new classifications.

**TABLE 3**

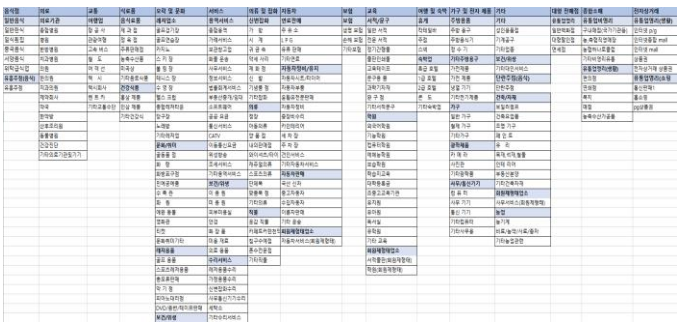


TABLE 3 details the classifications we have arbitrarily determined. The top row contains the new classifications we set, and the data below the column is placed with appropriate Middle and Small divisions from the original data column. In other words, the steeper color is new classifications, the one with brighter color is Middle division and white ones are Small division.

### *C. Data preprocessing*

The knowledge of making input data on N2D model is as follows: First of all, the new column, called new category name is added as a column on original data. Since the raw data has 100 million rows, it takes too much time and cost to apply the function we made with ‘Pandas’ library, so we changed the data type from ‘string’ into ‘int by mapping each unique words then by using ‘Numpy’ library we handled the data as vector which totally solved out the time costing problem.

After adding a column, we grouped the data by ‘date, region, gender, age and income’ in the original data. Therefore, the rows in the new data frame contain the aforementioned customers’ personal information, and the columns contain a new category. The values in the rows show the weight of which category a particular customer consumes, resulting in the sum of each row being 1. By grouping and dividing the data, approximately 100 million original data lengths were reduced to about 270,000 data lengths. This final compressed data is planned to go into input data in the N2D model.

## IV. APPLICATION

### *A. Training N2D model*

We tuned the hyperparameters empirically, not using hyperparameter tuning method such as gird search or Bayesian Optimization due to time cost and lack of computer environment. We set the number of clusters between 4 and 8. Then, we set the epochs which refer the single learning of the entire data as fixed value 30. And batch size, meaning the number of data belonging to on small group when training dataset is divided, was set the value as 32 or 64. Finally, unlike original n2d paper, which only utilizes the UMAP with two number of components, we set it as two or three. This is because we thought that we could find out better classified cluster not only with UMAP’s outcome dimension with 2 but also 3.

**Fig2**

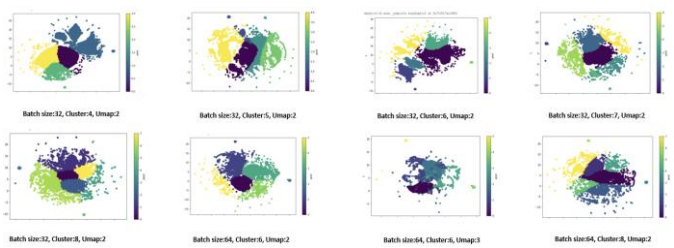



Fig2 shows image representation of the results of input data classification into a cluster, converting hyperparameter values one by one.

## V. Results and Discussion

### A. Measuring clustering performance with boxplot

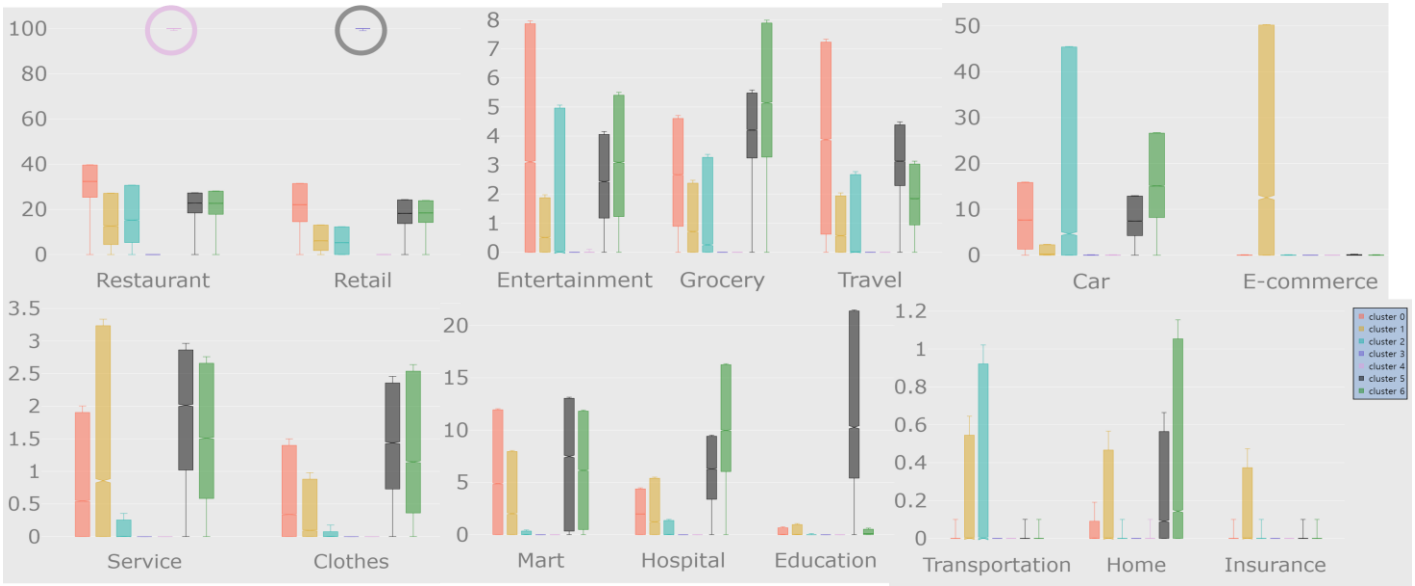
In this study, each cluster represent consumption pattern of BC card users. And clusters are characterized by categories such as ‘Restaurant’, ‘Retail’, ‘Car’, ‘Hospital’, and so on. A cluster’s characteristics are described by how much it spends on such categories. For example, if a cluster consists of group of users who spend most of their money on ‘Travel’ and ‘Entertainment, we can say that this cluster represents consumption on leisure activities.

For capturing cluster characteristics, boxplot is used because it shows distribution of each cluster on each category. Boxplot is a method for graphically depicting groups of numerical data through their quartiles. However, note that the figures of boxplot in this paper are presented without max value by capping the max with Q3(third quartile) + 0.1. This is because max values make boxplot too small to see its median, Q1(first quartile), and Q3(third quartile).

### B. Clustering results

Here, we present a clustering result that shows the best performances among results with different parameters. In other words, clustering characteristics appear clearly the most.

**Fig3**




In Fig3, boxplots indicate how much a cluster id distributed on each feature. For example, the green box (cluster_6) on the feature ‘Retail’ says that most people in this cluster spend around 20% of their total spending on ‘Retail. And a cluster’s characteristics are defined with the following two perspectives. First is ‘Which feature takes up a large portion of a cluster’s spending?’. For example, the red box (cluster_0) on ‘Restaurant’ says that the feature ‘Restaurant’ takes up around 30% of this cluster’s total spending. Second is ‘Which cluster shows a relatively high spending than others?’. For example, you can see on the feature ‘E-commerce’, the yellow box (cluster_1) dominates the other clusters. This cluster spends on ‘E-commerce’ than any other clusters.

Noticeably, there are two particular clusters that spend all their money in a single feature. As you can see on Fig3, cluster_4 (circled in pink) put 100% of spending in ‘Restaurant’ and cluster_3 (circled in grey) put 100% of spending in ‘Retail’. These clusters do not appear at any other features. They can be seen as a noise of raw data that is inevitable, because BC card does not solely represent one’s entire spending. It seems that some group of people used their BC card only when eating at restaurants or shopping at retail stores.

Consequently, according to boxplot, each cluster shows their representative consumptions like below. And here, clusters 3 and 4 are exceptional cases.

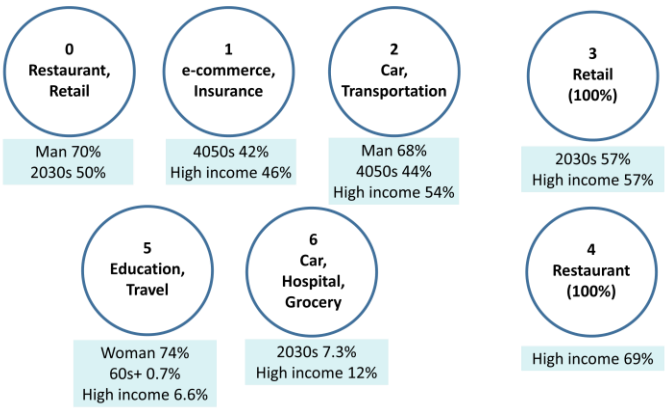
**Cluster 0: Restaurant, Retail**
**Cluster 1: e-commerce, Insurance**
**Cluster 2: Car, Transportation**
**Cluster 3: (Retail 100%)**
**Cluster 4: (Restaurant 100%)**
**Cluster 5: Education, Travel**
**Cluster 6: Car, Hospital, Grocery**

This clustering result was obtained from 3-dimensions, which had been reduced from its original dimension (16-dimensions). And clustering performance is better than when clustered in 2-dimensions because exceptional cases (cluster 3 and 4) that spend 100% in a single feature were not captured. Now, in further analysis, clusters will be examined and their characteristics will be discussed.

### C. *Further analysis – User characteristics*

To see what groups of people compose each cluster, user characteristics are examined. There are 7 clusters in total and proportions of users' demographic features are shown below figure.

**Fig4**




In Fig4, cluster 0 shows most typical consumption pattern (Restaurant, Retail) as it takes largest number of groups of people. Many users in 20s and 30s belong to this cluster. Cluster 1 and 2 show that people in 40s and 50s with high income have a consumption power on e-commerce, insurance, car, and transportation. In cluster 5, women in 20~50s dominate in spending for education and travel. In cluster 6, middle-aged and senior citizens dominate in spending for hospital.

### D. *Further analysis – Cluster transitions*

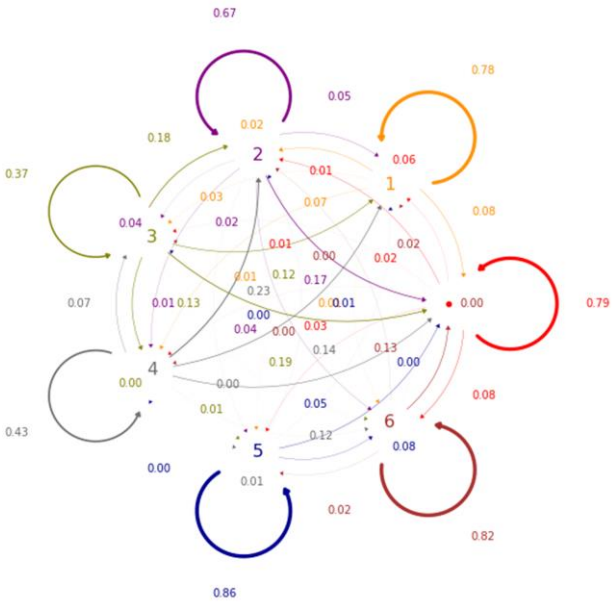
In this part, transitions between clusters are examined. Discrete-time Markov chain is used for producing one-step transition probabilities. States are defined as cluster belongings of group of users. There are 22,700 different user groups in pre-processed data which is a combination of gender (M/W), age (20/30/40/50/60+), income (B1~B11), and region (207 regions). Transitions are defined as month to month transitions. Fig5 below shows a part of transitions from Jan 2019 to April 2020, of different group of users. Numbers from 0 to 6 indicate cluster numbers.

**Fig5**


{'male_20s or less_B2_가평군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0],
'male_20s or less_B2_강남구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
'male_20s or less_B2_강동구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_강릉시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_강북구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_강서구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
'male_20s or less_B2_강진군': [5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_강화군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_거제시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_거창군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_경산시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_경주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_계룡시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_계양구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_고령군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_고성군': [0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_고양시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_고창군': [5, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0],
'male_20s or less_B2_고흥군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_공주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_과천시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_관악구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_광명시': [0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_광산구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_광양시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_광주시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_광진구': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_괴산군': [5, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_구례군': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_구로구': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
'male_20s or less_B2_구리시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_구미시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
'male_20s or less_B2_군산시': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],


Based on 249,511 transitions above, one-step transition diagram is produced for easier capturing of transitions. Fig6 below shows that cluster 3 and 4 have high probabilities to move to other clusters in next month. This result implies that if a customer spends entirely on either 'Retail' or 'Restaurant', it is likely that he will spend on other categories in the next month. In contrary, cluster 5 and 6 have high probabilities to stay in the same state in the next month. It implies that customers who mainly spend in 'Education', 'Travel', or 'Car', 'Hospital', 'Grocery' are likely to keep spending in the same categories in the next months.

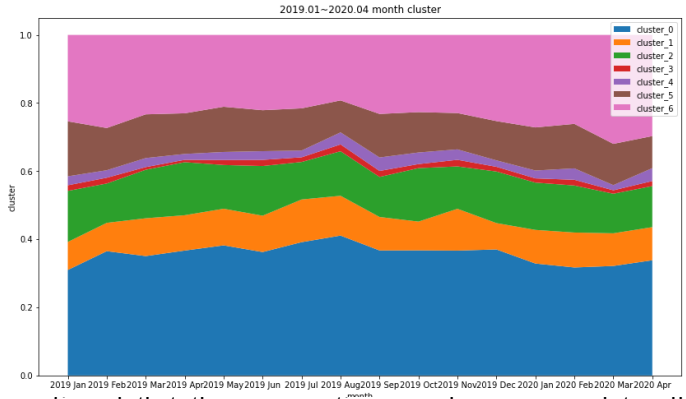
**Fig6**
**[one-step transition diagram]**



### *E. Further analysis – Monthly changes of clusters*

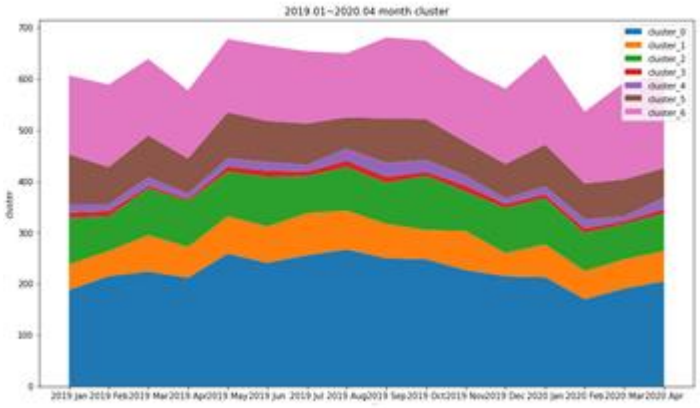
A graph was drawn to see monthly consumption patterns by looking at how the percentage of clusters change by month. The number of clusters was analyzed as 7, which show the overall graph flow well. The x-axis of the graph shows the values from January 2019 to April 2020, and the y-axis is the percentage value for each cluster divided by the sum of the number of clusters per month. A graph was presented, but there was no visible difference from month to month. Cluster ratio was almost the same for all months. This means that even if the seasons change, the consumption areas within the categories we set do not change that much.

**Fig7**




confirmed that the consumption was low compared to all periods, and the clusters also decreased all at once. It seems that the marked point is because the corona virus has spread a lot. In order to observe the change in detail at that time, the graph was drawn by gender and age.

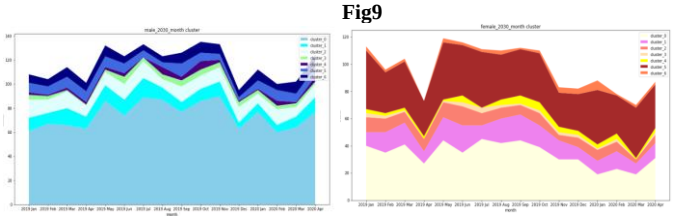
**Fig8**




### *F. Further analysis – Monthly changes of clusters by gender and age*

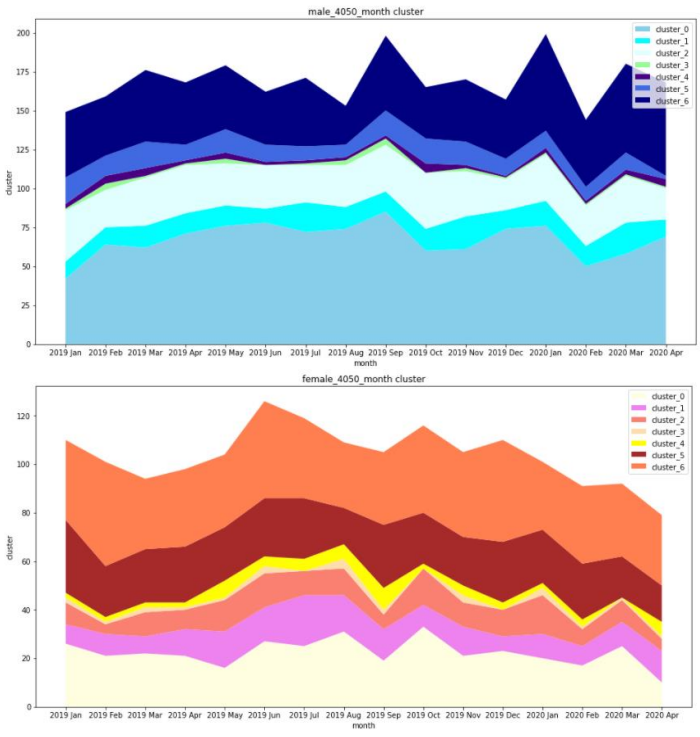
The shape of the monthly consumption change graph for men and women in their 20s and 30s was relatively similar. They saw a decrease in consumption in April 2019 and a sharp decrease in consumption during the Corona period in February-March 2020. Cluster_0, which has a high ratio in men, is characterized by restaurants and retail, and cluster_5, which is many in women, is characterized by travel and education, so it seems to have reacted sensitively to Corona.

**Fig9**



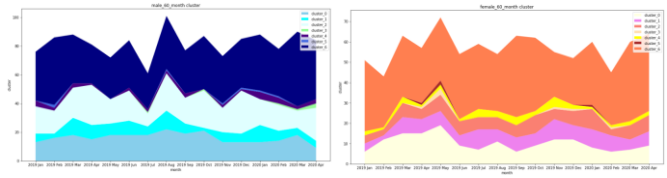
The graphs of men and women in their 40s and 50s were not as similar as those in their 20s and 30s. They also noticed that consumption decreased during the Corona period, and men quickly recovered to average consumption the next month, but women did not recover and continued to decline. It was clearly revealed that there is a difference in consumption between men and women in their 40s and 50s after experiencing the corona virus. Finding this difference in people in their 40s and 50s with strong consumption power seemed like useful information in the future.

**Fig10**




After the age of 60, the graphs were not similar to each other, and it seemed that there were individual differences. However, there are more clusters including hospitals than other age groups overall. It cannot be said that it has been greatly affected by the corona virus. From the age of 60, there seemed to be continuous consumption (like treatment cost in hospital) rather than sudden consumption.

**Fig11**



## VI. Conclusion

To summarize the project, we received BC card data from 2019.01 ~ 2020.04, did data cleansing, and clustered after processing such as category classification. In clustering, it was judged that the features were best shown when viewed as a boxplot by setting it to 7, and we wanted to obtain meaningful results through further analysis with this. First, the characteristics of users and consumption characteristics of the cluster were identified. Second, we looked at the transitions between clusters by month. By checking how consumption patterns change, we were able to obtain more useful information for predicting card usage. Finally, we looked at cluster usage by month, and we wanted to see what consumers were spending most on which month. Here, we could mainly infer information related to the corona, and it was clear that consumers had reacted to the corona.

And there were some limitations while working on the project. First, if we had more long-term data, we could have clearly identified the monthly rules. And in the case of a cluster consisting of 100% restaurant, the BC card did not show enough results to represent the overall consumption, and if there were other types of credit card data, it would have been more meaningful to analyze consumption. The second is about classification criteria. Unlike the N2D paper referenced in the project, there is no answer, so there is no label, so subjective interpretation is involved a lot. Even when creating 16 new categories, it is difficult to determine whether the subsets are properly distributed by randomly classifying them. Subjective opinions were also included in the interpretation of the graph, but it seemed to be the limit of not knowing the exact answer. So, follow up research simply mentioned that if there is more data, various patterns can be identified.

## References


[1] M. King, B. Zhu, and S. Tang, "Optimal path planning," *Mobile Robots*, vol. 8, no. 2, pp. 520-531, March 2001.

[2] McConville, Ryan, et al. "N2d: (not too) deep clustering via clustering the local manifold of an autoencoded embedding." arXiv preprint arXiv:1908.05968 (2019).