1. **데이터 정제**

주문별로 포함된 SKU의 총량과 유사도를 분석하기 위해 데이터 정제작업을 실시하였다. 6월부터 8월까지의 CJ 온마트 실제 주문데이터를 아래 FIGURE 1과 같이 일별, 차수별로 구분하고 각 차수에 포함된 주문 항목 수와 SKU 가짓수를 계산하여 나타내었다.

이어서 정리한 결과를 바탕으로 데이터 분석을 위해 샘플링을 진행하였다. 먼저, MPS에서는 하루 평균 10차수를 처리하므로 평균 차수 이하의 주문데이터 중 6월 5일 3차와 같이 주문 상품 수가 적은 데이터는 추출과정에서 손실 혹은 결손이 발생하였다고 판단하였다. 그리고 본 프로젝트 팀이 개발하고자 하는 알고리즘은 많은 주문을 실시간 처리해야 하므로 주문 상품이 많은 순으로 샘플링을 진행하였다. 데이터 샘플링 결과는 아래의 TABLE 1에 나타내었다.

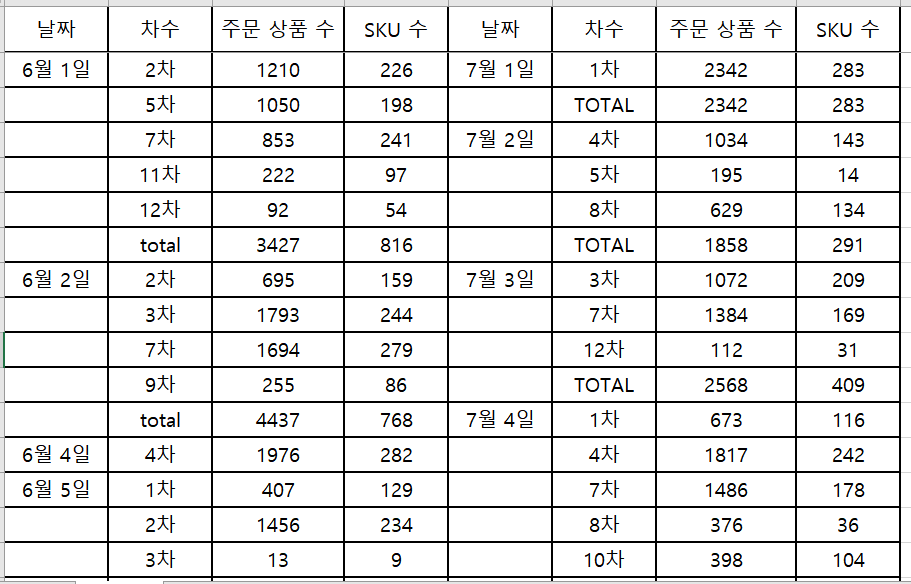
****

Figure 1 6, 7월 CJ 온마트 주문데이터 일별/차수별 데이터 정제

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 데이터 샘플링 결과 | | | | | | | |
| 6월 | | | | 7월 | | | |
| 6월 23일 | 6차수 | 4269 | 352 | 7월 20일 | 9차수 | 5028 | 390 |
| 6월 13일 | 6차수 | 3584 | 384 | 7월 6일 | 5차수 | 4282 | 262 |
| 6월 28일 | 5차수 | 3498 | 317 | 7월 23일 | 11차수 | 4267 | 368 |
| 6월 15일 | 5차수 | 3396 | 350 | 7월 19일 | 7차수 | 4175 | 392 |
| 6월 29일 | 9차수 | 3124 | 336 | 7월 26일 | 7차수 | 3325 | 343 |

Table 1 데이터 샘플링 결과

1. **데이터 분석**

이번 절에서는 1절에서 샘플링한 데이터를 가지고 유사도를 분석해보고자 한다. 유사도를 비교할 수 있는 알고리즘은 여러가지가 있으나 본 프로젝트팀은 K-MEANS 클러스터링과 연관 관계 분석을 사용하고자 한다. 각 샘플 데이터에 두 분석방법을 적용하여 분류 정확성이 높거나 배치 할당에 좀 더 용이한 결과값을 내는 분석방법을 택하여 프로젝트를 진행한다. 실제 구현하는 알고리즘에서는 차수 별 SKU 주문량이 일정수준을 넘어가거나, 5개 미만인 상품에 대해서는 분석을 진행하지 않으나 이번 절에서는 분석기법에 대한 판단이 주 목적이므로 차수에 포함된 모든 상품에 대한 분석을 진행한다. 또한 데이터 분석도구로는 R 3.4.2.version을 사용한다.

* 1. **K-MEANS 클러스터링**

클러스터링(군집 분석)이란 주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)를 정의하고 데이터 집단을 대표할 수 있는 대표점을 찾는 분석기법이다. K-MEANS는 그 중에서도 대표적인 분석기법이며 각 구간을 나눈 다음 중심점(Centroid)를 찾고 중심점을 기준으로 구간을 다시 나누고 변경사항이 있을 경우 다시 중심점을 찾아가는 과정을 반복하며 데이터를 군집화 한다. R을 이용하여 클러스터링을 진행하는 코드는 아래의 FIGURE 2와 같다.

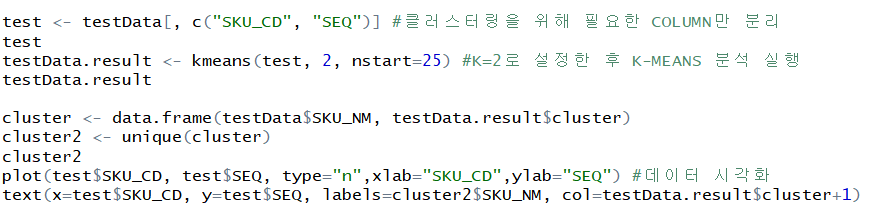


Figure 2 R을 이용한 K-MEANS 클러스터링

K-MEANS 클러스터링은 입력하는 변수와 함께 중심점의 갯수(군집의 수, K)를 지정하는 것이 중요하다. 이에 NbClust 패키지를 설치하고 FIGURE 3과 같이 NbClust 함수를 이용하면 FIGURE 4와 같은 도표를 생성할 수 있다. 이를 통해 Criteria값이 가장 높게 나오는 K를 파악하고, 클러스터링에 사용하여 최적의 결과를 도출할 수 있다.

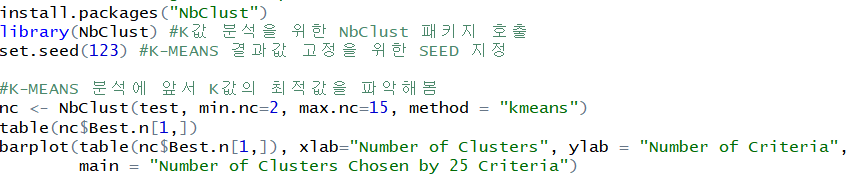


Figure 3 NbClust Package를 이용한 최적 K값 도출

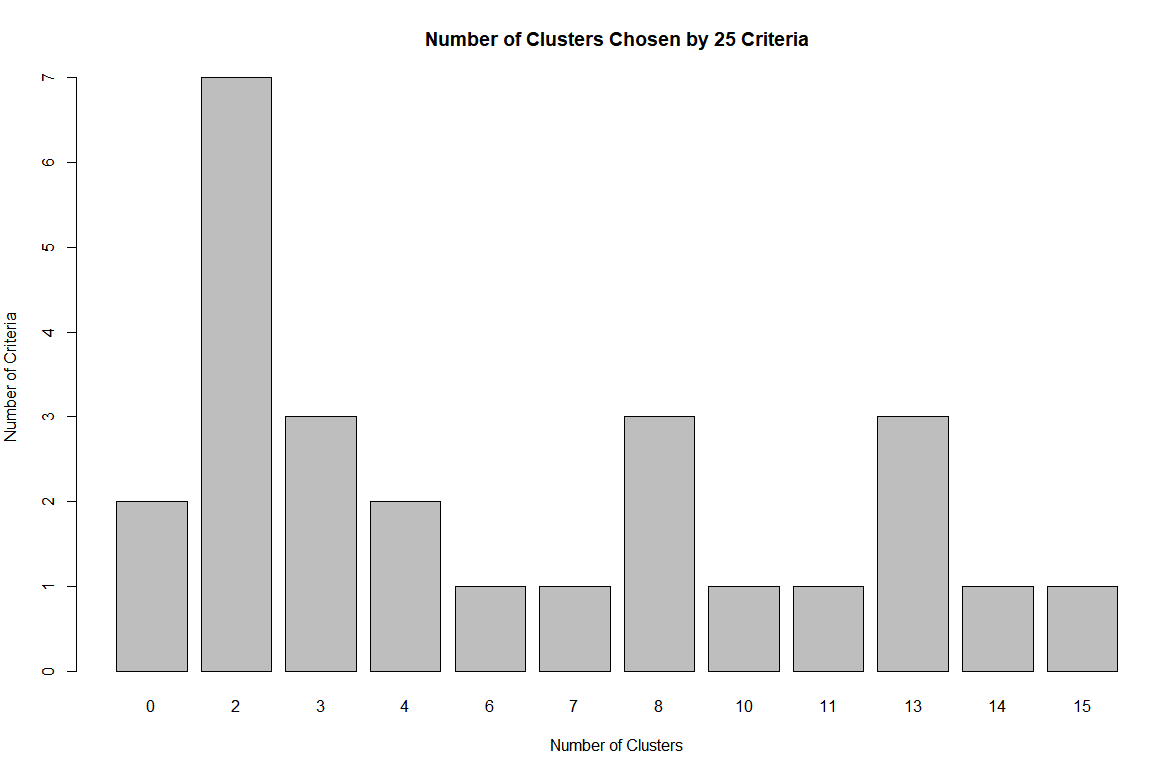


Figure 4 NbClust을 이용한 Barplot 출력

* 1. **연관 관계 분석**

연관관계 분석은 장바구니 분석이라고도 하며 고객의 대규모 거래데이터로부터 함께 구매가 발생하는 규칙을 도출하여, 고객이 특정 상품 구매 시 연관성 높은 상품을 추천하는데 사용되는 분석기법이다. 평가척도로는 지지도, 신뢰도, 향상도를 사용하며 이에 대한 세부적 설명은 TABLE 2와 같고, R에서는 arules 패키지를 이용하여 FIGURE 5와 같이 모델링 하였다.

|  |  |  |
| --- | --- | --- |
| 지지도 | 전체 거래 횟수에서 특정한 두 항목 A,B가 거래된 횟수가 차지하는 비율 | A,B 동시 거래 수 / 전체 거래 수 |
| 신뢰도 | A를 산 사람들 중에서 A와 B를 동시에 산 사람들의 비율 | A,B 동시 거래 수/ A 거래 수 |
| 향상도 | 구매 고객 중 B를 산 사람이 차지하는 비율과 신뢰도의 비율 | 신뢰도 / 전체 거래에서 B가 포함된 거래가 차지하는 비율 |

Table 2 연관 관계 분석 평가척도

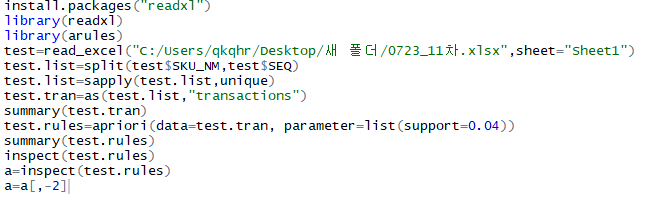


Figure 5 R을 이용한 연관 관계 분석

* 1. **분석기법 평가**

10개의 샘플링 데이터에 대한 클러스터링과 연관 관계 분석을 실시하였고, 그 중 한 사례인 6월 29일 9차수 DATA에 대한 분석을 예시로 분석기법의 평가를 진행하고자 한다.

먼저 클러스터링의 경우 NbClust 함수를 이용하여 K값을 탐색하였고, 아래의 FIGURE 6과 같은 결과가 나왔다. 이를 토대로 중심점의 개수 K를 2로 설정하고 SKU\_CD가 6자리 이상인 사은품 항목 20개을 제외하여 316항목에 대해 K-MEANS 분석을 진행하였다. 분석 결과로 첫번째 클러스터에는 228가지 SKU, 총 2187개의 주문이 포함되었고, 두번째 클러스터에는 88가지 SKU, 819개의 주문이 포함되었다. 이를 FIGURE 7과 같이 시각화하여 나타내었다.

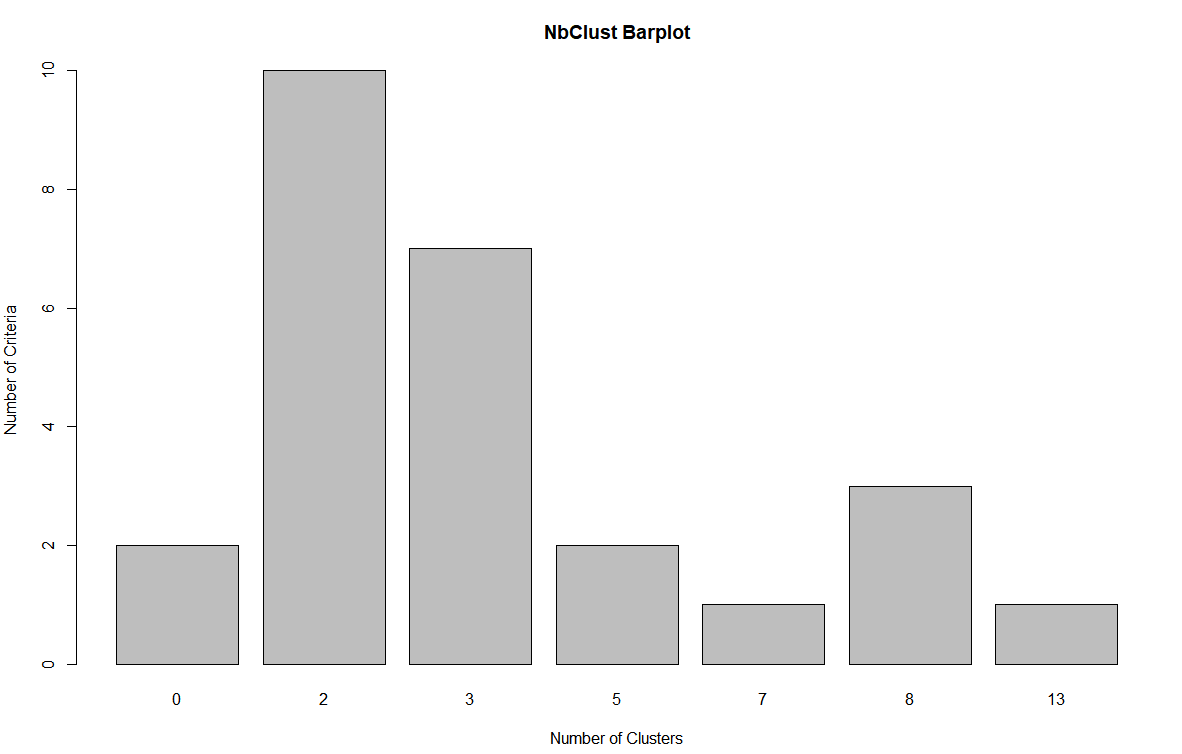


Figure 6 0629\_9차수 NbClust 분석 시각화

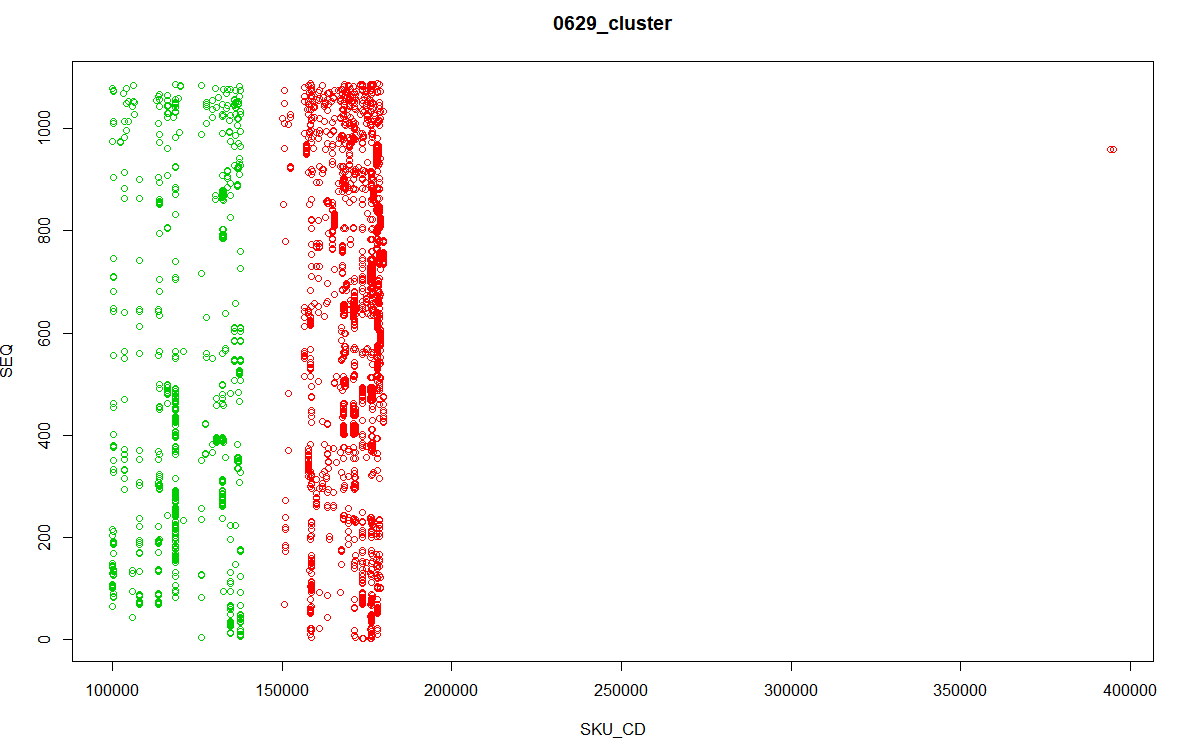


Figure 7 0629\_9차수 K-MEANS 클러스터링 시각화

다음으로 연관 관계 분석결과를 살펴보자. 아래의 FIGURE 8과 같이 분석 결과가 나왔으며, 가장 많이 포함된 상품은 비비고/육개장 500G (NS)/상온, 비비고/사골곰탕 500G (NS)/상온, 스팸마일드 200g, 스팸클래식 200g/상온, 비비고/미역국 500g/상온임을 알 수 있다. 이 차수에서는 상품 조합의 지지도 기준을 0.023으로 설정했고, 8개의 조합이 추출되었으며 FIGURE 9와 같다. 가장 많이 포함된 조합은 {햇반/컵반 사골곰탕국밥 166G/상온, 햇반컵반 미역국밥 167G/상온}이고, 8개의 조합이 비슷한 횟수로 추출되었다.

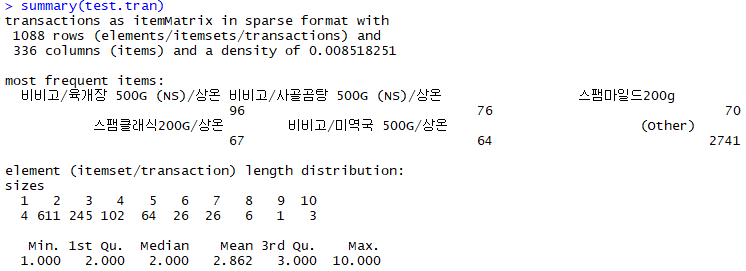


Figure 8 0629\_9차수 연관 관계 분석



Figure 9 0629\_9차수 추출된 조합

위 결과들을 종합해 보았을 때, 클러스터링은 섹션 단위 별 SKU 구분만 가능하지만 연관 관계 분석은 추출된 조합을 토대로 배치 할당이 가능하므로 프로젝트의 목적에 더 적절하다고 판단할 수 있다. 또한 클러스터링의 경우 SKU\_CD와 SEQ를 기준으로 분석을 진행하였는데, 클러스터를 구성하기에 기준이 부족하여 중심점의 개수가 적게 나오고 균일하게 나뉘지 않는다는 단점이 있으므로 연관 관계 분석을 활용하여 프로젝트를 진행하겠다.