클러스터링을 이용한
대학생 카페 선호 분석 및
비즈니스 적용방안

과목명: 경영과 정보 시스템 월567
담당교수: 강현정 교수님
제출자: 2조
B431217 이준건
B598105 표병수
B631195 이은비
B631286 한서현

1. 서론

(1) 주제 선정

클러스터링 과제 주제로 대학생의 카페 이용 **행태(?)를** 선정하였다. 대학생을 대상으로 고려해보았을 때, 카페가 실생활에 밀접한 소재로 익숙한 주제이기 사실적이고, 신빙성이 높은 데이터를 수집하기에 용이하기 때문이다. 이 데이터를 통해서 대학가 주변에서 카페를 개업하고자 하는 이나, 현재 카페를 운영 중이지만 운영상 리모델링이 필요한 사람들에게 비즈니스적 시사점을 제공하고 비즈니스에 적용할 수 있게 하고자 한다.

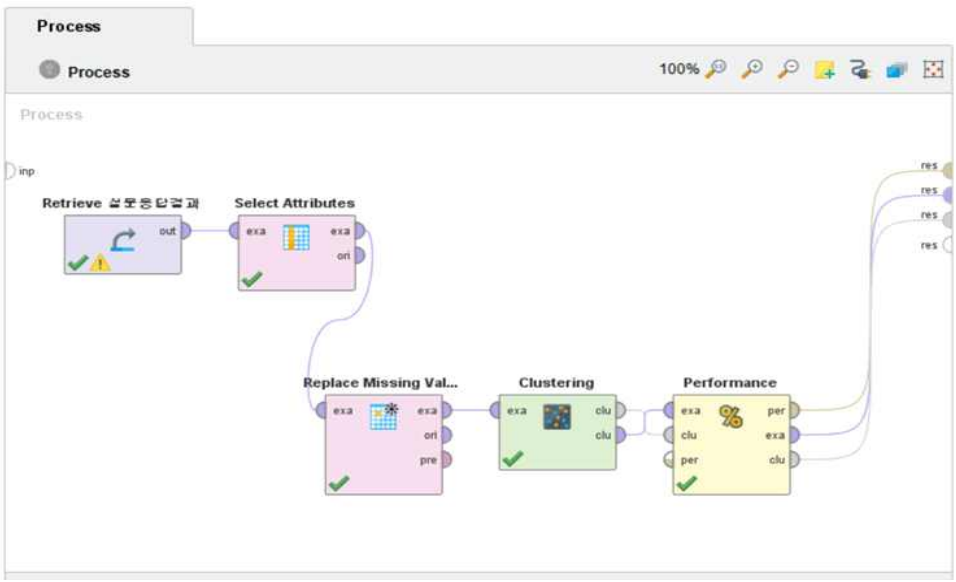
(2) 설문지 작성 및 데이터 수집.

설문지의 항목은 비즈니스에 활용하기 위해 사용자(ex, 점주) 입장에서 소비자의 특성을 분류할 수 있는 항목과 대학생이 카페를 선택할 때 중요하게 생각하는 변수들이 무엇일까 크게 이 두 가지를 고려하여 구성하였다. 그 결과 많은 변수가 있지만, 방문 빈도수, 가격 민감도, 분위기 민감도, 평균 체류시간, 프랜차이즈 선호도, SNS 민감도, 메뉴 다양성, 선호 매장 규모, 주 이용 시간대, 프로모션 민감도 총 10개의 항목으로 설문지를 구성하였다.

설문지의 경로는 접근성이 높고, 사용이 쉬우며, 엑셀 파일 화가 용이한 구글폼을 선정하였다. 데이터 수집 대상이 대학생이므로 대학생의 자료만 수집하기 위해, 설문지는 경영과 정보시스템 과목을 수강하는 학생들과 팀원 각자의 대학생 지인들로 구성된 채팅방에 공유하였다. 그 결과 최종적으로 56개의 데이터를 수집 할 수 있었다.

2. 본론

(1) 프로세스 모델링



클러스터링을 수행하기 위해 위한 프로세스는 다음과 같이 구성하였다. (1) CVS 형식으로 저장된 설문 조사 자료를 가져온다. 이때, 타임라인은 필요 없는 데이터이므로 Exclude column을 선택하여 보이지 않게 설정한다. (2) 클러스터링에 사용할 변수의 종류를 설정하기 위해 Select Attributes 오퍼레이터를 연결해준다. 이때 필터 타입은 subset을 선택하여서 여러 변수를 넣을 수 있게끔 설정해 준다. (3) 혹시 모를 missing values로 인한 오류를 없애기 위하여 Replace missing values 오퍼레이터를 연결해준다. (4) K-means Clustering 오퍼레이터를 연결한다. 이때, 자료를 5점 척도로 수집하였기 때문에 measure type은 Numerical Measure로 설정을 해주고, Numerical measure는 Euclidean Distance로 설정을 해주었다. K-means Clustering의 Parameters를 설정할 때, 가장 고민하였던 부분은 k 즉, 군집 개수와 max run이었다. 이 두 가지를 어떻게 설정하느냐에 따라서 클러스터링이 어떻게 형성되는가가 달라지기 때문이다. 어떻게 하면 가장 합당한 군집을 찾아낼 수 있을까를 다방면으로 고민한 결과, 군집 개수를 3으로 설정하고 max run은 1000으로, max optimization steps는 1000으로 설정하였다. (5) Clustering Distance Performance 오퍼레이터를 연결하여서 main criterion은 Davies Bouldin으로 설정을 하였다.

(2) 클러스터링 분석 및 비즈니스 적용

위 과정을 통해 프로그램을 실행해 본 결과 유의미한 클러스터링을 몇 가지 찾아내었다.

3. 결론

(1) 기타 개선점 및 마무리

설문 조사 결과를 데이터화하고, 이 데이터를 군집화해보니 설문지에 포함되어있는 10가지 속성 중에 어떤 것은 큰 의미를 가지지만, 굳이 설문지에 넣지 않았어도 상관이 없었을 것 같은 속성들도 발견하게 되었다. 미리 알았다면 그 질문들 대신 다른 질문을 하나 더 넣었으면 좀 더 확실한 클러스터링이 되지 않았을까 싶은 생각도 든다. 또 클러스터링을 돌려보니, 질문지를 만들 때 왜 상관도가 높은 질문들로 구성을 하라고 하셨는지 이해를 할 수 있게 되었다.

또 이번 클러스터링에 설문 조사를 통해 얻은 데이터는 이전에 실습했던 iris 데이터와 달리 명확하게 군집이 형성되고, 답이 존재하는 데이터가 아니기 때문에 데이터를 분석하기가 쉽지만은 않았다. 이 과정에서 특히 군집 개수와 max run을 어떻게 설정하면 Davies Bouldin 수치와 Average-within-centroid distance를 줄일 수 있을까 즉, 어떻게 하면 군집의 응집력을 높일 수 있을까를 고민하였고, Parameters의 속성을 여러 번 바꾸어 보았다. 그 결과 완벽하지는 않더라도 나름대로 가장 가시성이 높고, 유의미한 클러스터링 결과를 얻을 수 있는 방안을 찾게 되었다.

위와 같은 과정을 반복해 본 결과, 이번에는 변수를 두 개만 넣고 클러스터링을 돌리는 것으로 결정하였다. 하지만 변수를 3개 4개 더 넣어서도 유의미한 결과를 발견할 수 있다면 그렇게 하는 것이 클러스터링의 본 의미에 조금 더 합당한 방법이 아니었을까 하는 아쉬움도 남는다. 그렇지만, 2개의 변수로도 충분히 인사이트를 발견하였고, 3개, 4개, 10개를 시도해보았을 때 사용할 만한 클러스터링이 구성되지 않기 때문에 최종적으로 위와 같이 구성하게 되었다.

분석 프로그램을 이번 학기에 처음 접해보고, 클러스터링을 처음 해보는 4명이 모여서 기초부터 하나하나 해보며 프로젝트를 수행한 결과, 물론 상기 서술한 어려움도 많이 존재하였다. 하지만 이러한 일련의 경험들을 바탕으로, 이번 프로젝트 이후에 클러스터링 분석을 할 기회가 있다면 이와 같은 문제점들을 보완하고 더 확실한 인사이트를 찾아내며 실제 비즈니스에 적용 가능한 분석을 시도할 수 있을 것이다.