**연관규칙 마이닝(Association rule mining)**

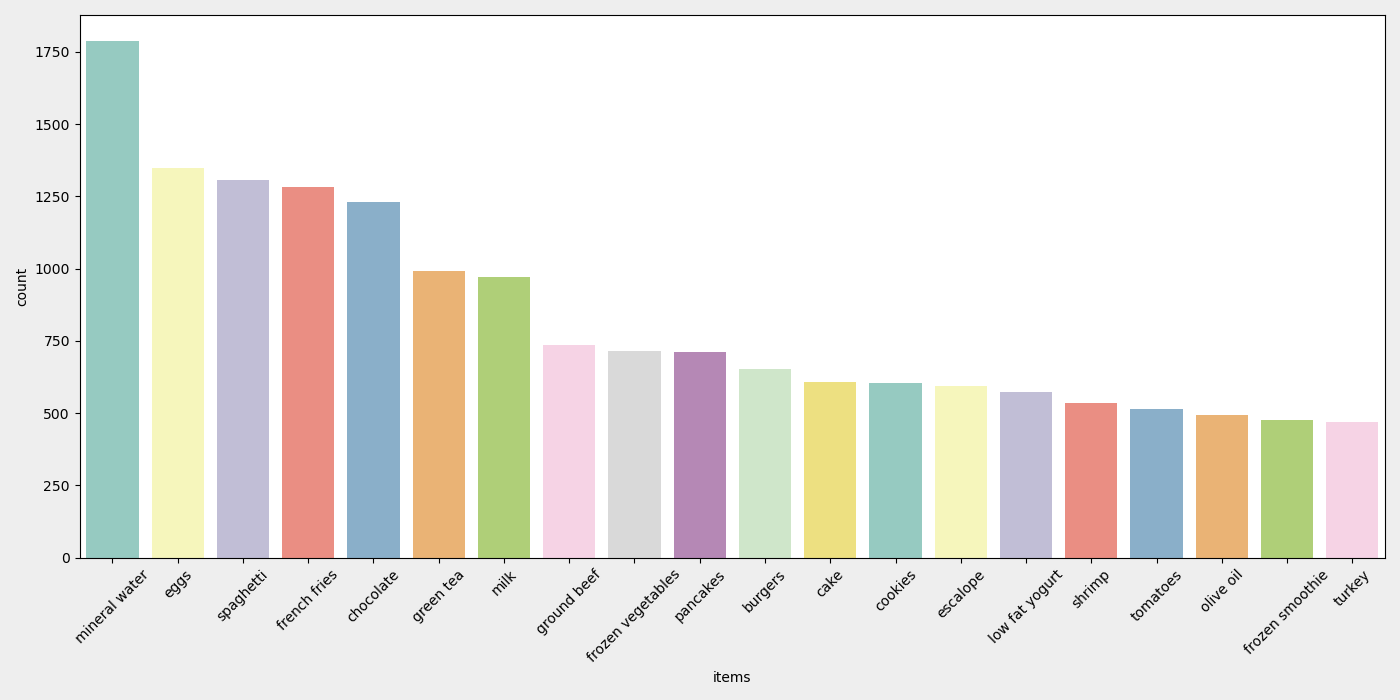
**- Apriori 알고리즘 -**

**결과 보고서**

202012252 지수영

1. **데이터**
2. Store\_data

* 장바구니에 많이 담긴 아이템 순으로 상위 20개 정렬



* 장바구니 수 : 7501
* 아이템 수 : 120 ('mineral water', 'eggs', 'spaghetti', 'french fries', 'chocolate', 'bramble', 'cream', 'napkins', 'water spray', ' asparagus' 등)

1. **Apriori 알고리즘 진행 순서**

(1) Min\_threshold 값 보다 큰 1-Frequent itemset을 찾는다.

(2) 1-Frequent itemset을 조합해 크기가 2인 2-itemset을 만들고 2-itemset에 있는

Item 중 1-frequent itemset에 없는 item을 제거하는 Prune 과정을 진행한다...

(3) 2-itemset에서 2-Frequent itemset을 찾는다.

(4) k-itemset의 item이 남지 않을 때까지 위의 과정을 반복한다.

1. **지표**

* 전체 경우의 수에서 집합 X와 Y를 모두 구매하는 경우의 수를 나타내는 지표
* 집합 X를 구매하는 경우에서 Y도 구매하는 경우의 비율

1. **Support threshold 변화에 따른 연관규칙이 달라지는 정도**
2. Min\_support

* support threshold 값에서 total(총 데이터(장바구니) 개수)로 나눈 값으로 설정한다.
* 제공된 데이터 Store\_data의 총 transaction 개수는 7501개로 **total = 7501** 이 된다.

1. Min\_confidence

* Confidence가 Min\_confidence 이상일 경우에만 연관규칙을 출력한다.

1. Support threshold 값 = 100

* 결과 1

1. ('mineral water', 'eggs', 'spaghetti')

Support : 0.014264764698040262

confidence ('mineral water', 'eggs') -> ['spaghetti'] : 0.2801047120418848

confidence ('mineral water', 'spaghetti') -> ['eggs'] : 0.23883928571428573

confidence ('eggs', 'spaghetti') -> ['mineral water'] : 0.3905109489051095

* {‘mineral water’, ‘egg’}를 구매했을 때 {‘spaghetti’}를 구매하는 경우가 있다.
* {'mineral water', 'spaghetti'}를 구매했을 때 {‘egg’}를 구매하는 경우가 있다.
* {‘eggs’, ‘spaghetti’}를 구매했을 때 {‘mineral water’}을 구매하는 경우가 있다.

1. Support threshold 값 = 200

* 결과 1
  1. ('green tea', 'mineral water')

Support : 0.03106252499666711

confidence green tea -> ['mineral water'] : 0.23511604439959638

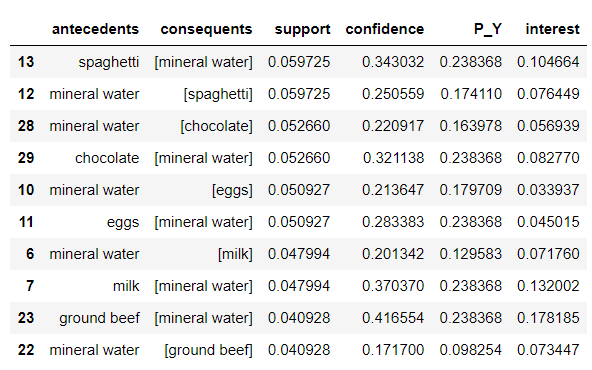
confidence mineral water -> ['green tea'] : 0.1303131991051454

* + Green tea를 샀을 때 mineral water를 구매하거나 mineral water를 샀을 때 green tea를 구매하는 경우가 있다.

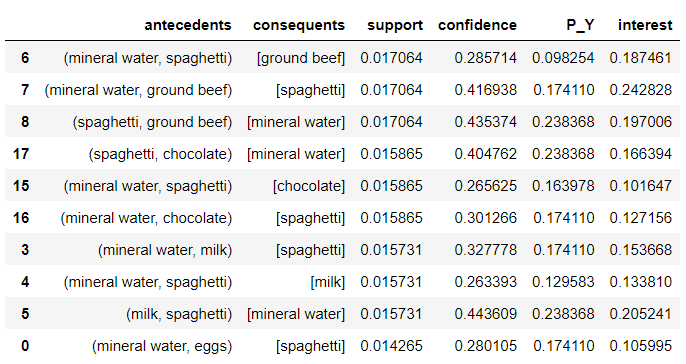
1. 결론

* Support threshold 값이 커질수록 Support의 X 아이템 수가 적어지는 것을 볼 수 있었다..

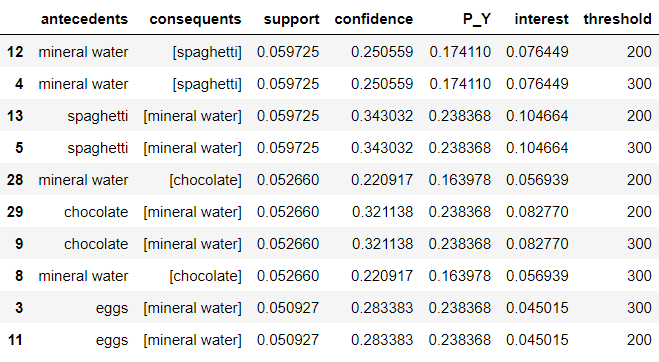
1. **정리**
   1. Support 값 기준으로 정렬(support threshold : 200)



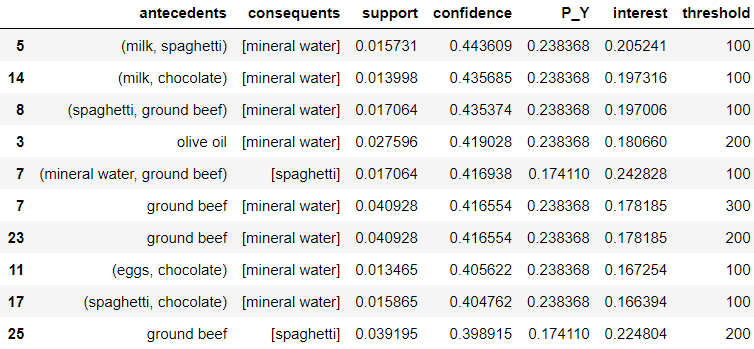
* 1. Support 값 기준으로 정렬(support threshold : 100)



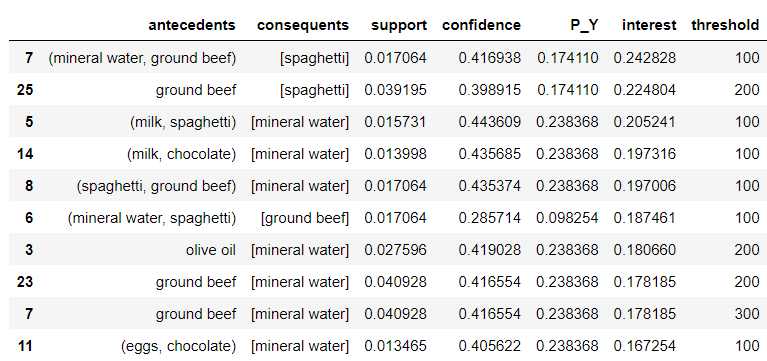
* 1. Min Threshold 값이 100, 200, 300일 경우를 모두 합친 후, support 값을 기준으로 내림차순으로 정리한 결과



* 1. Min Threshold 값이 100, 200, 300일 경우를 모두 합친 후, confidence 값을 기준으로 내림차순으로 정리한 결과



* 1. Min Threshold 값이 100, 200, 300일 경우를 모두 합친 후, interest 값을 기준으로 내림차순으로 정리한 결과



**[결론]**

* Threshold 값이 클수록 antecedents 집합의 원소 개수는 줄어들고, support 값은 커지는 경향을 보였다.
* 반면 Threshold 값이 작아질수록 anteedents 집합의 원소 개수는 늘어나고, confidence 값은 커지는 경향을 보였다.
* Interest 값을 기준으로 result(세 threshold 값으로 도출된 결과를 합친 것)를 정렬했을 때, Min Threshold 값이 작을수록 상위에 위치해 있었다.