프로그래머스 최종프로젝트 보고서

클러스터 적용 후 사용할 df

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그 중 사용할 칼럼들 선정 ( 고객ID / 제품카테고리 / 수량  / 배송료 / 쿠폰상태 / 구매금액 / 초중말 / 월 / 요일 /  성별 /  고객지역 / 오프라인비용 / 온라인비용 / Cluster ) + count를 위한 거래횟수 칼럼추가

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

분석해볼 두 데이터프레임 ( Premium / Churn )

텍스트, 흑백, 문서, 화이트이(가) 표시된 사진

자동 생성된 설명

커다란 함수를 만들어서 해당 데이터프레임 안에서 칼럼끼리의 상관관계를 분석할 수 있도록 만들었다 ( GetInfo(dataframe, column) ).

함수에 데이터프레임, 타겟칼럼을 집어넣으면 GetInfo( df, col )  
1. 판매수량, 거래횟수, 구매금액의 합계값을 가진 데이터프레임

2. 구매금액, 배송료, 오프라인비용, 온라인비용의 평균값을 가진 데이터프레임

3. 타겟칼럼의 고유 고객 수에 대한 데이터프레임

4. 타겟칼럼의 쿠폰상태 값(sum)을 가진 데이터프레임

5. 각 제품카테고리별, 타겟칼럼의 매출, 판매수를 나타낸 데이터프레임

6. 시간대별 타겟칼럼의 구매금액 값(sum)을 가진 데이터프레임

7. 요일별 타겟칼럼의 거래횟수에 대한 데이터프레임

등이 있다

각 데이터프레임들을 통해 알 수 있는 인사이트들을 찾아보자

데이터 셋은 Premium 고객들로, 방문율도 높고 자주 구매하고, 높은 실적을 남기는 고객들이다.

일단 비교적 간단한 데이터프레임(1)을 예시로 들어 타겟칼럼에 대한 것을 설명하고 앞으로의 데이터프레임들 (2~7)에 대해서는 한가지 예시 (고객지역)를 들면서 결과를 설명하도록 할 것이다.

1. 판매수량, 거래횟수, 구매금액의 합계값

해당 데이터프레임은 그래프보단 숫자로 보는 것이 편할 것 같아 단순 데이터프레임만 출력하도록 했다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명( target column = 성별 )

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명( target column = 고객지역 )

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명( target column = 초중말 )

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명( target column = 월 )

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명( target column = 성별 )

텍스트, 스크린샷, 번호, 메뉴이(가) 표시된 사진

자동 생성된 설명( target column = 성별 )

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명( target column = 성별 )

인덱스에 타겟 칼럼이 들어가고, 그에 따른 데이터들을 볼 수 있다. 이런 방식은 타겟을 바꿔가며 분석을 할 수 있다. 지금은 위에서 예고한대로 ‘고객지역’ 관련해서 분석을 하겠다.

Chicago쪽(중서부)에서 거래가 가장 많았고 [55059], 그 다음은 California(서부) [40642] New Jersey [11583] + New York [23074] (북동부), 마지막으로 Washington DC (남부) [5731]로 판매수량 순위가 매겨졌다. 한 번의 거래에서 판매되는 수량도 위의 순서대로였지만 [ 4.8 > 4.6 > 4.0 > 3.6 ] 오히려 판매된 물품 중에서 판매된 물건의 평균가는 역순을 기록했다 [ 19.6 < 21.6 < 22.4 < 25.0 ]

여기서 확실히 거래가 많이 이루어지는 지역 (Chicago)에서는 한번 거래할 때 많은 양을 하다보니 물건을 저렴하게 구매하는 경향이 있었다

1. 구매금액, 배송료, 오프라인비용, 온라인비용의 평균값

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

여기서의 ‘구매금액’은 총 금액을 거래횟수로 나눈 값으로, 위에서 계산한 물건당 평균가격과 다르다. 즉 Chicago를 예시로 들면 한번의 ‘거래’에서는 평균 94.2를 지불했지만, 하나의 ‘물건’ 값은 19.6을 지불한 것이다. 배송료를 봤을 때, 물류창고에서부터 거리가 멀면 배송비가 좀 더 많이 나갈 것이라고 가정해보면 물류창고의 위치도 대략적으로 추정이 가능할 것이다. ( Ohio로 생각 ) California는 온라인비용이 적은데 비해 오프라인비용은 크고 Chicago는 전부 적절하게 비싸고, New Jersey는 모든 비용이 적다. New York은 California랑 반대로 오프라인비용보단 온라인비용이 많이 나갔다. Washington DC는 딱 중간이다. 따라서 California / New York 에서는 온/오프라인으로 광고하면 효과적으로 손님을 끌어들일 수 있고, 나머지 지역들은 판매방식에 따른 이득을 보진 못할 것 같다.

1. 타겟칼럼의 고유 고객수

해당텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명 지역에서 거래한 고유 유저 수도 거래량에 비례하지만, 고객 1명당 사용한 평균 금액은 각각 [ 8129, 10095, 7603, 5917, 9570 ]로 New York에서는 평균적으로 적은 사용금액을 가졌고 Washington DC도 눈에 띄게 높은 수치를 나타냈다. 이는 New York은 거래하는 사람이 많아서, Premium 고객층에서 그나마 낮은 분들이 계시고, Washington DC지역은 그 반대로 몇몇의 whale 유저들이 지역의 매출을 담당하고 있다고 볼 수 있다.

1. 타겟칼럼의 쿠폰상태

이젠 그래프를 통해 알아보자

도표, 그래프, 라인, 텍스트이(가) 표시된 사진

자동 생성된 설명

쿠폰사용에 따른 거래횟수 그래프이다.

모든 지역이 비슷한 양상을 보이고 있다 ( Clicked > Used > Not Used ) 일단 지금 단계에서는 더 이상 발견할 인사이트는 없는 것 같다.

1. 타겟칼럼의 각 제품카테고리에 대한 판매수,매출

5-1. 판매수

텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

지역별 제품 카테고리에 대한 판매 수 그래프이다. 가장 많은 판매량을 가진 top 5개의 데이터만을 사용해서 그래프로 나타내었다. 모든 지역에 공통적으로 Apparel, Nest-USA, Office를 많이 구매하였고, Nest, Drinkware, Lifestyle에 관한 소비가 지역마다 4~5등을 다르게 차지했다.

5-2. 매출

텍스트, 도표, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

도표, 그래프, 텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

지역별 제품카테고리에 대한 매출 그래프이다. 특이하게도 가장 많은 판매량을 기록한 Apparel이 매출에 관해서는 3위를 차지하고 Nest-USA가 판매량은 2위, 매출은 압도적 1위인 것을 볼 수 있었다. 매출 2위는 Nest로 해당 제품의 판매량이 상위 5개 안에 들지 않아도 매출로 상위권을 찍을 만큼 이익을 많이 낸다고 볼 수 있다. ( 적게 팔아도 매출이 높음 ) 그 외에는 Drinkware 또한 판매량 상위권에 속한만큼 매출도 상위권을 담당하는데, New Jersey에서는 Bags가 Drinkware대신 자리를 차지했다.

1. 타겟칼럼의 각 시간대에 대한 구매금액
   1. 월별

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

지역별 월별 매출 추이다. Chicago나 California은 시간의 흐름에 따라 매출량이 증가했고, 5월달(May)에 한풀 꺾였다가 다시 증가했다. 전체적으로 봤을 때 11월( Nov )에서 매출이 높으므로 해당 기간에 마케팅을 진행하면 좋을 것 같다. 추가적인 정보들은 딱히 알 수 없으므로 다른 타겟칼럼을 활용해야 할 것 같다. ( 제품카테고리 )

6-2. 초중말

여기는 1일~10일을 상순(초) 11일~20일을 중순(중) 21일 이후를 하순(말)로 나눠서 분기별로 판매량을 확인하기 위해 추가로 만든 칼럼을 활용했다. 해당 사진은 1월 달을 확인했다.

라인, 도표, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명

California는 시기에 상관없이 비슷한 소비를 하는 것으로 보이고 Chicago와 Washington DC는 중순에 상대적으로 많은 소비가 이루어지는 것을 볼 수 있다. New York과 New Jersey는 특이하게도 반대 성향을 가지는데 이는 점점 소비가 증가하고 감소하는 것이다. 물론 1월만 봐서 확답은 할 수 없고 다음 데이터들도 봐야 한다.

1. 타겟칼럼의 각 요일에 대한 거래횟수

라인, 도표, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명

지역별 요일별 판매량이다. 지역마다 다르지만 대부분 수요일~금요일 사이에 많은 판매가 이루어지므로 해당 기간 동안 마케팅을 하여 소비를 유도하는 전략을 세울 수 있을 것이다.

같은 방식으로 Premium 🡪 Churn으로 데이터셋을 바꿔서 진행했다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

1. .

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그래프, 도표, 라인, 직사각형이(가) 표시된 사진

자동 생성된 설명

1. .

5-1.

텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

스크린샷, 텍스트, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

5-2.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명 텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

스크린샷, 텍스트, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

1. .

6-1.

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명 텍스트, 도표, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명

6-2.

라인, 도표, 텍스트, 그래프이(가) 표시된 사진

자동 생성된 설명

1. .

라인, 도표, 그래프, 경사이(가) 표시된 사진

자동 생성된 설명

한계

아무래도 모든 데이터의 합산값을 가져가므로 유저별 맞춤 추천이 어렵다.

클러스터가 너무 골고루라서 차별점이 크진 않을 것 같다. (고유 고객 수?)

칼럼을 통일시켜야 할듯하다. 거래횟수 <> 매출 ( 그냥 둘 다 쓰자 )

도메인 지식이 부족하여 해석에 한계를 느낀다