거래와 구매 행동 간의 관계: 거래량과 상품 주문량 분석

- StudentID: 22100331

- Name: 방은주

- 1st Major: 상담심리

- 2nd Major: 데이터사이언스

이 분석에서는 날짜별 입출금 활동과 고객의 상품 주문 건수 간의 관계를 탐구해 볼 것이다. 이를 위해 BT 데이터와 S 데이터를 분석하여 입출금 활동과 상품 주문 건수 간의 상관관계를 조사할 것이다. 입출금 활동은 고객의 금융 거래 활동을 나타내며, 상품 주문 건수는 고객의 구매 행태를 나타낸다. 이 연구를 통해 날짜별 거래량 정도가 고객의 상품 주문 건수에 미치는 영향을 파악하고, 금융 거래와 구매 행태 간의 관계를 이해하는 데 기여할 것으로 기대된다.

1. Data overview

데이터는 각각의 CSV 파일에서 검색하고 R을 사용하여 사전 처리하였다. 날짜는 적절하게 서식을 지정하고 결측치 값은 0으로 대체하였다. 특히, 이 분석은 2020년 1월 1일부터 2022년 12월 31일 사이의 기간으로 제한되었다.

1.1 BT 데이터

샘플 크기: 5,000,000개
변수: 5개
변수 유형: 모두 문자형
데이터 범위: 2020년 8월 21일 ~ 2155년 11월 15일

1.2 S 데이터

샘플 크기: 5,000,000개
변수: 14개
변수 유형: 숫자형과 문자형
데이터 범위: 2010년 1월 22일 ~ 2020년 12월 2일

1.3 데이터 분석 방법

분석에는 dplyr, ggplot2 와 같은 패키지와 R 프로그래밍 언어를 사용했다. BT 데이터는 제한된 날짜에 대하여 요일과 월별 거래량을 계산하는 데 사용되었다. 거래량과 주문량을 시각화하기 위해 BT와 S 데이터 세트에서 고유한 항목을 추출했다. 이 추출한 항목을 바탕으로 데이터를 합쳐 제한된 날짜 안에서 거래량과 주문량과의 관계를 분석한다. 이를 분석하기 위해 군집 분석과 상관분석을 사용하였다.

2. Univariate analysis

입출금이라는 변수에 초점을 맞추어 특정 요일이나 월에 많은 거래가 일어나는지 확인하려고 한다.

2.1 요일별 거래 분석

데이터에서 요일을 추출하여 각 요일별로 입출금 거래가 어떻게 분포되어 있는지 확인한다.
요일별로 입출금 거래의 평균 또는 총합을 계산하여 시각화하고, 어떤 요일에 가장 많은 거래가 발생하는지 확인한다. 단, 기간을 2020 ~ 2022년 사이로 한정한다.

	weekday_transaction	Average_Transaction
Friday	12129	12023.71
Monday	12373	12023.71
Saturday	11529	12023.71
Sunday	12694	12023.71
Thursday	10358	12023.71
Tuesday	13550	12023.71
Wednesday	11533	12023.71

Figure 1. 요일별 거래량

위 표를 살펴보면 2020에서 2022년간 요일별 평균 거래 금액이 12,023.71으로 나타난다. 이 평균 거래 금액에 각 요일별 평균 거래 금액을 비교했을 때 많은 거래가 이루어진 요일은 화요일이다. 가장 적은 거래가 이루어진 요일은 목요일로 확인 가능하다.

2.2 월별 거래 분석

데이터에서 월을 추출하여 각 월별로 입출금 거래가 어떻게 분포되어 있는지 확인한다. 월별로 입출금 거래를 입금과 출금으로 나누어 계산하여 시각화하여 어떤 월에 가장 많은 거래가 발생하는지 확인한다. 이를 막대형 그래프를 이용해 월별 거래량을 사용한다.
이때, 기간을 2020에서 2022년 사이로 한정하여 계산한다.

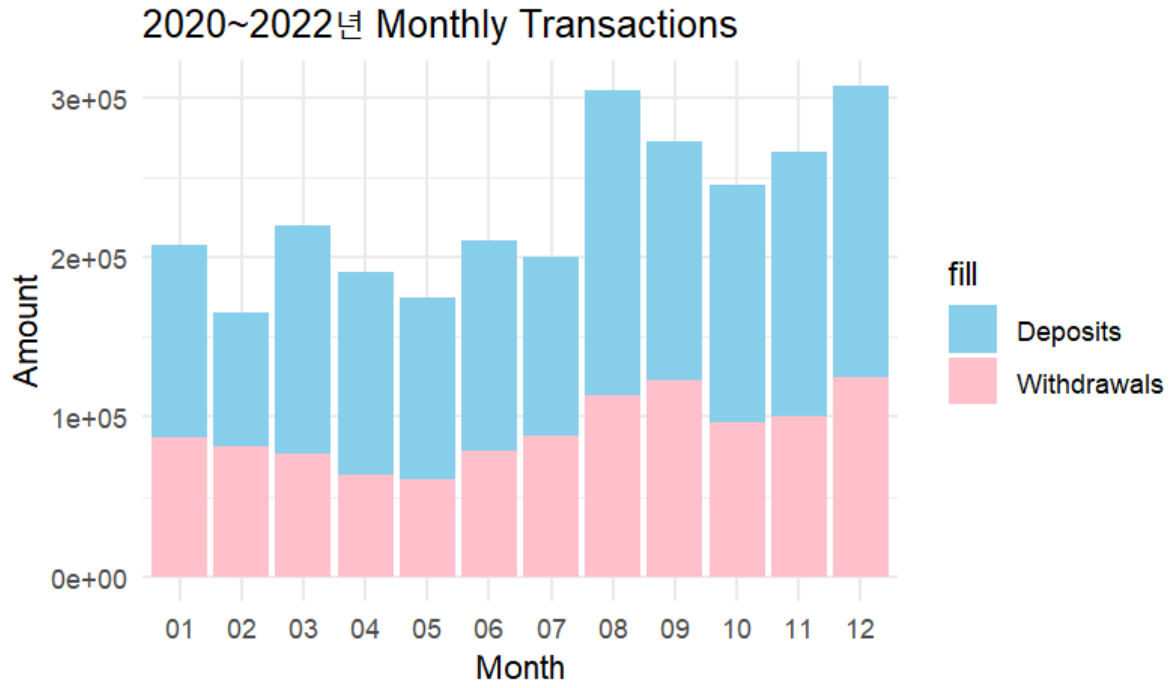

Figure 2. 2020~2022년 Monthly Transactions

2020년부터 2022년까지의 월별 거래 데이터를 나타내는 이 차트는 예금과 출금 거래의 패턴을 비교 분석한다. Y 축은 거래 금액을 나타내며, 0에서 300,000 사이의 값으로 표시된다. 이 값들은 '1e+05'와 같이 표현된다. X축은 1월에서 12월까지의 월을 나타내며, 각 월은 두 종류의 거래를 나타내는 막대그래프로 구분된다.
그래프를 살펴보면 연말에 입금 거래가 크게 증가하는 경향이 보인다. 특히, 10월과 11월은 다른 달들에 비해 입금 거래 활동이 활발한 것으로 나타났다. 반면, 출금 거래는 전반적으로 더 균일한 분포를 보였지만, 봄철인 3월과 4월에 소폭의 증가가 있음을 보인다.

3. Multivariate analysis

상관관계 및 클러스터링 기법을 사용하여 변수 간의 숨겨진 패턴을 탐색하여 거래량과 주문량에 대한 이해를 제공한다. 이를 수행하기 위해 날짜에 기반하여 BT 데이터와 S 데이터를 합쳐서 2020에서 2022년에 해당하는 데이터로 분석을 진행하였다.

3.1 Clustering

클러스터링 분석, 특히 K-평균을 사용하여 거래 행태에 따라 데이터를 별개의 그룹으로 분류했다. 데이터는 세 그룹으로 클러스터링 되어 서로 다른 기간에 해당할 수 있는 패턴을 드러내도록 만들었다.
아래 그래프에서 'Total Orders'를 x축으로, 'Total Deposits'를 y축으로 한 산점도는 이미 이 두 변수 간의 관계를 시각적으로 나타내고 있다.

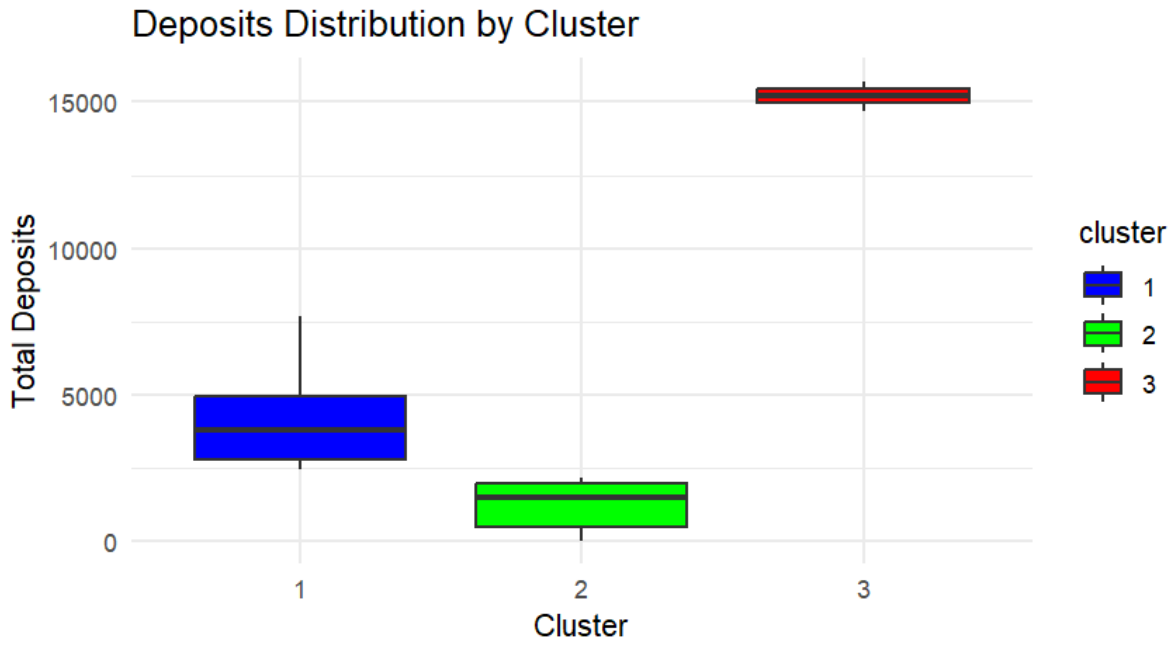

Figure 3. 거래량과 클러스터 간의 관계

3.1.1 클러스터 센터

클러스터 중심값은 다음과 같다:

total_orders
  클러스터 1: 537.0000
  클러스터 2: 536.9091
  클러스터 3: 536.8750

total_deposits
  클러스터 1: 15200.945
  클러스터 2: 1249.938
  클러스터 3: 4161.289

total_withdrawals
  클러스터 1: 5647.5450
  클러스터 2: 288.8509
  클러스터 3: 1954.5850

3.1.2 클러스터 분석

클러스터 중심 값을 분석하여 클러스터별 평균 주문량과 평균 총 거래량을 비교할 수 있다. 클러스터 1은 매우 높은 총 거래량과 상대적으로 높은 주문량을 가지는 데이터 포인트들로 구성되어 있다. 이는 높은 주문량과 높은 거래량이 관련 있을 수 있음을 시사한다. 클러스터 2는 낮은 거래량과 주문량이 비교적 비슷한 데이터 포인트들로 구성되어 있다. 이는 주문량이 일정 수준에서 증가하더라도 총 거래량의 증가가 크지 않음을 나타낼 수 있다. 클러스터 3은 중간 정도의 총 거래량과 주문량을 가진 데이터 포인트들을 포함한다. 이는 클러스터에서 주문량과 총 거래량 사이에 어느 정도의 상관관계가 있다는 것을 나타낸다.

3.2 Correlation

상관관계 분석은 총 주문량과 총 거래량 간의 관계의 강도와 방향을 파악하기 위해 수행하였다. 변수 간의 상관관계를 측정하는 피어슨 상관관계 계수를 계산하였다.

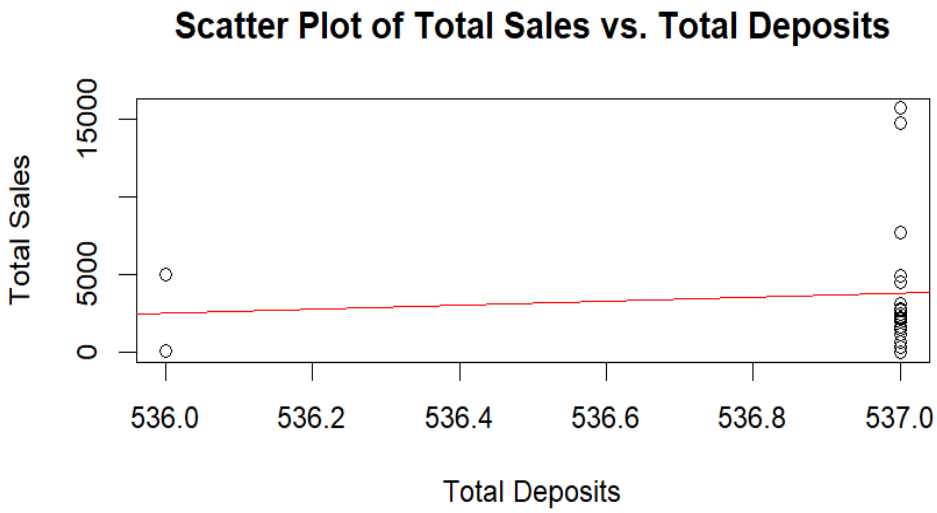

Figure 4. 2020~2022년 주문량과 거래량 상관관계

피어슨 상관관계 계수의 계산된 결과는 `0.09080288`로 나타났다. 이는 총 주문량과 총 거래량 사이에 상관관계가 약하게 나타나는 것을 알 수 있다. 이를 통해 거래량이 주문량에 영향을 미칠 수는 있으나 다른 요인이 더 작용할 수 있다는 것을 의미한다.

4. Suggestion

단변량, 클러스터링, 상관관계 분석을 포함한 BT 및 S 데이터 세트에 대한 탐색적 데이터 분석을 통해 고객의 거래량과 주문 행동 간의 관계에 대해 얻을 수 있었다. 특정 요일이나 월에 거래량이 증가하는 것을 알 수 있었다. 이를 통해 판매하는 곳에서는 해당 요일에 할인이나 주문량을 증가시키는 전략을 고려할 수 있다.
총 주문량과 거래량 간의 상관관계는 약했지만 S 데이터에 해당하는 다양한 구매 패턴을 가진 그룹들은 발견할 수 있었다. 이는 매출에 가장 크게 기여하는 클러스터를 파악하여 그들 개인에 맞는 마케팅 전략을 개발할 수 있음을 의미한다.

<프로젝트>

거래 및 주문 데이터를 통해 식별된 행동 클러스터를 기반으로 타겟팅 된 판매 전략을 개발하는 것을 목표로 한다. 이 접근 방식을 통해 고객 구매 행동에 대한 미묘한 차이를 파악하여 개인화된 마케팅 전략을 만드는 것이 가능할 것이다. 더 나아가 매출에 가장 크게 기여하는 클러스터를 파악하고 이들의 특정 니즈에 맞는 서비스나 상품을 개발할 수 있을 것이다.

목표: 마케팅의 개인화&니즈에 맞는 제품 개발&행동 트렌드 분석

데이터 활용: BT 및 S 데이터 세트를 더욱 활용하여 추가적인 행동 지표와 인구통계학적 데이터를 통합하여 클러스터링 모델을 개선한다. 거래 행동과 구매 패턴에 영향을 미칠 수 있는 외부 경제 요인을 포함하도록 분석을 확장한다. 머신러닝 모델을 구현하여 과거 거래 데이터를 기반으로 미래의 주문 행동을 예측하여 이를 마케팅 전략에 적용한다.

기여: 고객 세분화를 강화하여 타깃 마케팅과 영업 전략의 효과를 높일 수 있다. 개인화된 상품 제공이 가능하여 고객 만족도나 유지율이 향상될 수 있다.

즉, 이 프로젝트는 금융 거래량과 구매 결정 사이의 연관성에 집중하여 기업이 고객의 요구를 충족하여 맞춤형 서비스를 제공하는 것을 목표로 한다. 이를 통해 전반적인 고객 만족도를 향상시키는 것이 궁극적인 이 프로젝트의 기여이다.