실습과제 2

빅데이터 6기

이신영

요약

쇼핑몰 데이터를 활용하여 분석을 진행하였다.

먼저, 전체 매출액을 기준으로 상위 3개 그룹인 다우기술, 지니, 천재 태블릿을 대상으로 RFM 분석을 실시하였다. R, F, M 분석에서 모두 다우기술이 가장 높은 값을 보였고, 지니와 천재 태블릿이 그 뒤를 이었다. 이를 바탕으로 입점 기업별 판매 전략 수립을 세울 수 있다. 다음으로, 월별 매출액은 3, 5, 10월이 높게 나타났고, 연도별 매출액은 2019년부터 2022년까지 꾸준히 증가하였다. 이를 통해 월별, 연도별 판매 전략을 수립할 수 있다. 마지막으로, 결제방법과 결제금액의 연관성을 분석한 결과, 신용 거래가 현금 거래보다 거래액이 높음을 확인하였다.

본 분석 결과는 지니마켓의 마케팅 전략을 수립하는 근거로 활용될 수 있을 것이다.

주요 용어

1. RFM 분석

   고객이 얼마나 최근에 구매를 했는가를 나타내는 최근성(R: Recency), 얼마나 자주 구매했는가를 나타내는 빈도(F: Frequency), 얼마나 구매했는가를 나타내는 구매액(M: Monetary Amount) 등 3개 요인을 분석하여 정량적으로 고객을 분류하는 방법이다.[1]

   이 보고서는 입점 기업을 대상으로 RFM 분석을 진행하므로 '얼마나 최근에 판매했는가', '얼마나 자주 판매했는가', '얼마나 판매했는가'를 분석하였다.

2. 초도상품

   개발이 완료된 제품에 대하여 실용 시제품과 동질의 물품이 생산되는가를 확인하기 위

[1] 박광호. (2002). 인터넷 소매유통업의 RFM 모델 기반 충성고객관리를 위한 웹서비스 (WsLCM) 프레임웍 (Web services Framework for Loyal Customer Management based on RFM Models in Internet Retailing). 지능정보연구, 8(1), 43.

하여 양산에 앞서 처음으로 소량 생산해 보는 제품이다.[2]

3. 매출

기업이 영업을 목적으로 하는 상품 등의 판매 또는 용역의 제공을 행하고 대가를 받음으로써 실현되는 수익(收益)을 말한다.[3]

4. 순수익

총 이익 중에서 영업비·잡비 등 총비용을 빼고 남은 순전한 이익이다.[4]

---

[2] 국방과학기술용어사전

[3] 조세통람

[4] 패션용어사전

목차

I 서론

1. 연구 주제 및 목적
2. 연구 질문 및 연구 가설

가. 연구 질문

나. 연구 가설

II 연구 방법

1. 데이터 수집 방법
2. 조사 방법 및 사용된 통계적 방법

III 결과

IV 토의

1. 결과 요약
2. 한계점
3. 제언

V 참고문헌

Ⅰ 서론

1. 연구 주제 및 목적

입점 기업별 RFM 분석의 목적은 다음과 같다.

첫째, 입점 기업을 최근성, 빈도, 금액 측면에서 세분화하여 각 요소에 맞게 타깃 마케팅 전략을 구성할 수 있다. 둘째, 우수 입점 기업을 식별한 뒤 해당 기업에게 특별 혜택이나 프로모션을 제공하여 고객 충성도를 높이고 이익을 극대화할 수 있다. 셋째, 입점 기업을 세분화함으로써 해당 기업에 적합한 다양한 마케팅 전략을 수립할 수 있다. 넷째, 최근에 판매를 하지 않은 기업을 식별하여 이탈을 방지하고 재구매를 유도할 수 있다.

종합하면, 본 연구에서는 RFM 분석을 통해 입점 기업을 세분화하고 각 그룹에 맞는 전략을 수립하여 고객 경험을 향상시키고 이익을 극대화할 수 있는 시사점을 제공하는 데 목적이 있다.

2. 연구 질문 및 연구 가설

가. 연구 질문

[필수 분석]
1) 입점 기업별 RFM 분석(3그룹)
2) 매출 시각화
가) 월별, 연도별 매출
나) 월별 순수익 (처리 상황, 할부기간 고려)
3) 결제방법에 따른 분석
가) 결제방법과 결제금액의 연관성 분석
나) 결제방법은 맨 앞의 한가지만 사용한 것으로 간주
[선택 분석]
1) 매출 시각화 2
가) 최대 매출 상품 3종류 집계
나) 주문 연도에 따른 해당 상품의 매출 증감 분석
2) 연관성 분석
가) 주문한 달과 판매금액의 상관관계 분석
※ 연도의 변화는 무시한다

나. 연구 가설

본 연구에서는 [필수 3]에 대한 가설을 설정하였다.

가설: 신용 거래가 현금 거래보다 거래금액이 클 것이다.

귀무가설(H0): 신용 거래와 현금 거래의 거래금액은 같거나 차이가 없다.
대립가설(H1): 신용 거래의 평균 거래금액은 현금 거래의 평균 거래금액보다 크다.

II 연구 방법

1. 데이터 수집 방법

본 연구는 천재교육의 '프로젝트 기반 빅데이터 서비스 개발자 양성 과정'에서 본사로부터 제공 받은 '미니프로젝트-쇼핑몰 실습데이터.xlsx' 데이터를 활용하였다. 데이터 정보는 다음과 같다.

- 데이터 기간
: 2019-01-01 ~ 2022-11-08

- 열 항목
주문번호, 업체명, 상품명, 제조사, 주문수량, 판매금액, 결제방법, 주문일자, 처리상태, 초도상품, 제작문구, 할부기간

2. 조사 방법 및 사용된 통계적 방법

가. 전처리

업체명, 상품명, 결제방법 열에서 결측치가 있는 행을 삭제하였다. 업체명 열의 '지니', '지니 태블릿', '지니 태블릿(후불집행)'은 같은 업체이므로 업체명을 '지니'로 통일하였다. 주문 수량과 판매금액 열의 값이 0인 경우, 해당 행을 삭제하였다. 제작 문구 내역 열은 분석 관련 내용이 없고 대부분이 결측치이므로 삭제하였다. 할부기간 열의 필드값이 없는 부분은 일시불 결제로 판단하고 '일시불'로 채웠다. 여러 결제방법이 사용된 경우, 첫 번째 결제방법만 사용한 것으로 간주하였다.

나. 입점 기업별 RFM 분석(3그룹)

총 89개의 입점 기업명을 기준으로 분류하고 이 중 총 판매금액이 가장 큰 3개의 입점 기업을 대상으로 RFM 분석을 진행하였다. R분석은 주문일자를 기준으로 2022년, 2021년, 2019~2020년으로 나누어 각 3점, 2점, 1점을 부여하여 계산하였다. F분석은 주문량 100이상, 50이상, 50미만으로 나누어 각 3점, 2점, 1점을 부여하여 계산하였다. M분석은 지출 금액 분포 75% 이상, 50% 이상, 50% 미만으로 나누어 각 3점, 2점, 1점을 부여하여 계산하였다.

다. 매출 시각화

1) 월별, 연도별 매출

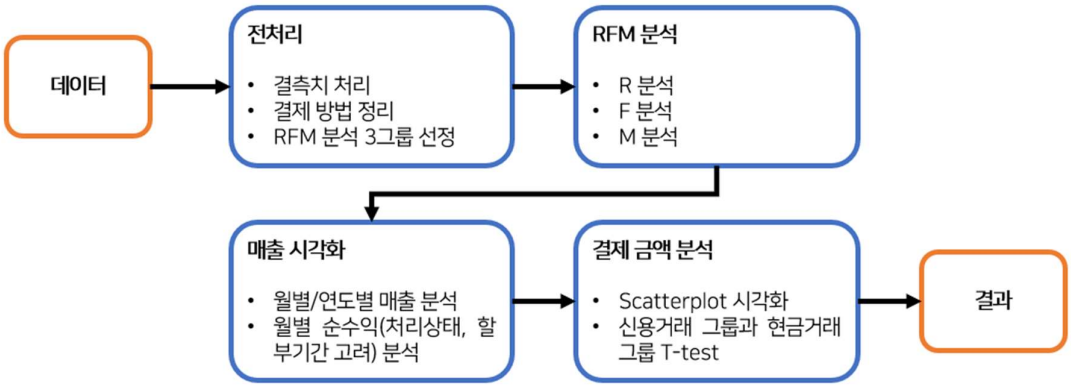
주문일자를 각각 월별, 연도별로 추출하여 월과 연도를 기준으로 그룹화한 판매금액을 계산하였다.

2) 월별 순수익(처리 상황, 할부기간 고려)

'처리상태'가 '구매확정'인 행의 판매금액만을 고려하여 순수익을 계산하였다. 이때, '할부기간' 열의 값인 '일시불', '1개월', '6개월', '12개월', '18개월', '24개월'를 고려하여 매월 순수익에 포함하여 계산하였다.

라. 결제방법에 따른 분석: 결제방법과 결제금액의 연관성 분석

결제방법은 맨 앞의 한 가지만 사용한 것으로 간주하였다. 신용카드, 정기결제, 후불을 신용 거래로 분류하고 웰컴마일, 포인트, 현금간편결제, 적립금, 가상계좌는 현금 거래로 분류하여 결제방법별 결제금액을 시각화 하였다. 검정 방법으로는 Mann-Whitney U Test를 사용하였다.

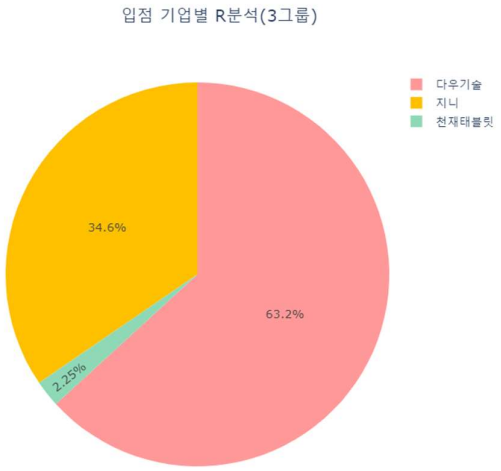


[그림 1. Workflow]

III 결과

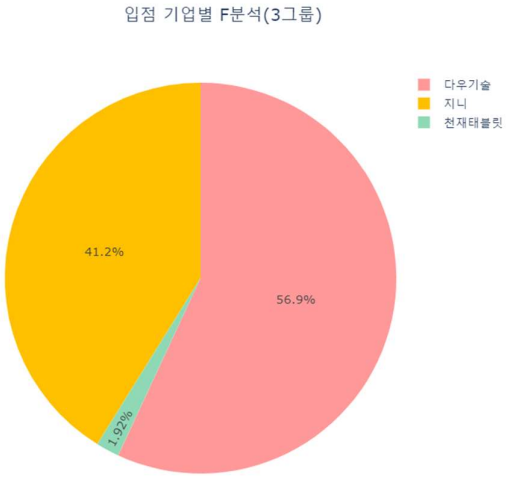
1. 입점 기업별 RFM 분석(3그룹)

R값 결과는 다우기술(63.2%), 지니(34.6%), 천재 태블릿(2.25%)로, 다우기술이 압도적으로 최근 판매가 많았다. 천재 태블릿은 2순위 지니와 비교했을 때 약 32%의 큰 차이가 났다.



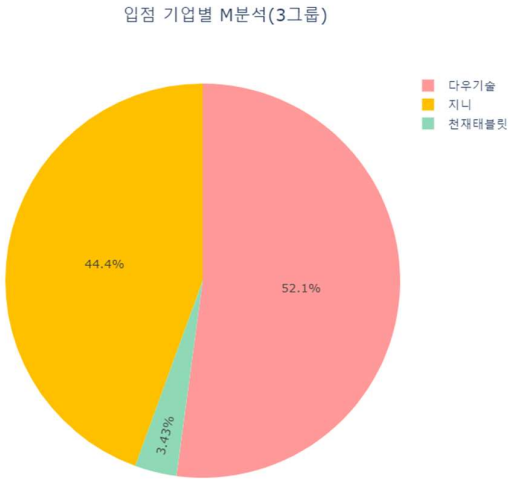
[그림 2. 입점 기업별 R분석(3그룹)]

F값 결과는 다우기술(56.9%), 지니(41.2%), 천재 태블릿(1.92%)로, 다우기술이 가장 자주 판매하였다. 천재 태블릿은 2순위 지니와 비교했을 때 약 39%의 큰 차이가 났다.



[그림 3. 입점 기업별 F분석(3그룹)]

M값 결과는 다우기술(52.1%), 지니(44.4%), 천재 태블릿(3.43%)로, 다우기술이 가장 많이 판매하였다. 지니는 1위인 다우기술과 약 8% 차이가 났는데, R분석과 F분석 결과에서 보다 다우기술과 적게 차이가 나는 것을 알 수 있다. 천재 태블릿은 2순위 지니와 비교했을 때 약 40%의 큰 차이가 났다.

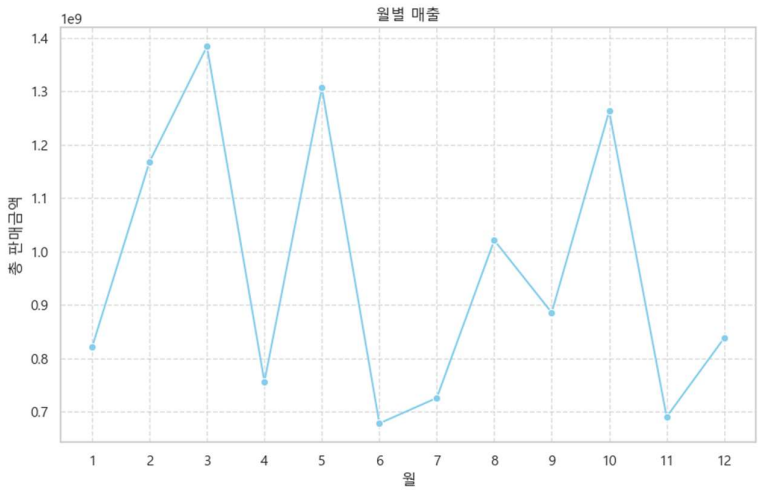


[그림 4. 입점 기업별 M분석(3그룹)]

2. 매출 시각화

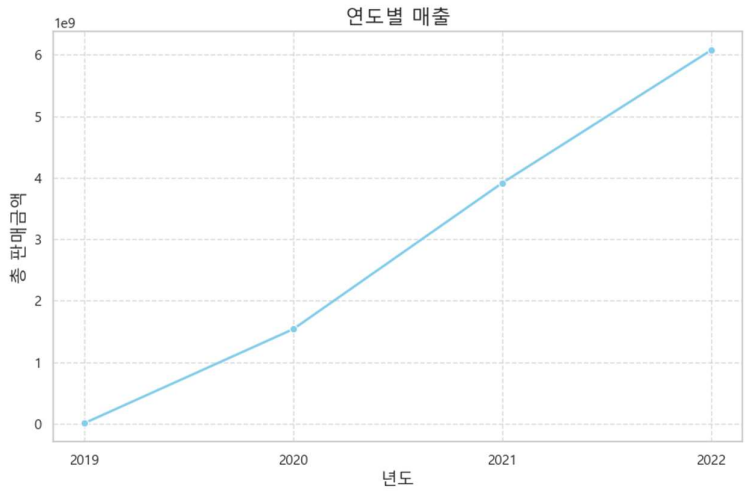
가. 월별, 연도별 매출

신학기가 시작되는 3월의 매출이 가장 크고, 차례로 5월과 10월 순으로 크다.



[그림 5. 월별 매출]

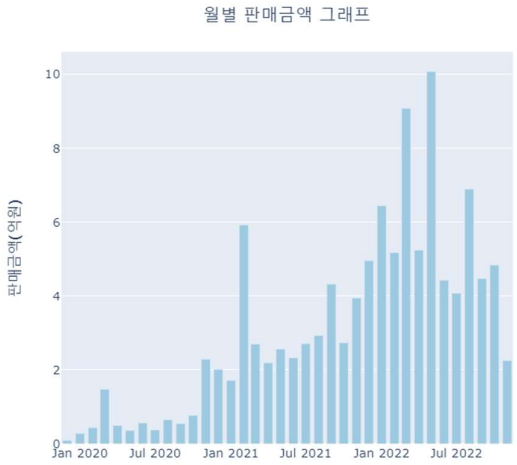
2019년부터 2022년까지 꾸준히 매출이 증가했다.



[그림 6. 연도별 매출]

나. 월별 순수익(처리 상황, 할부기간 고려)

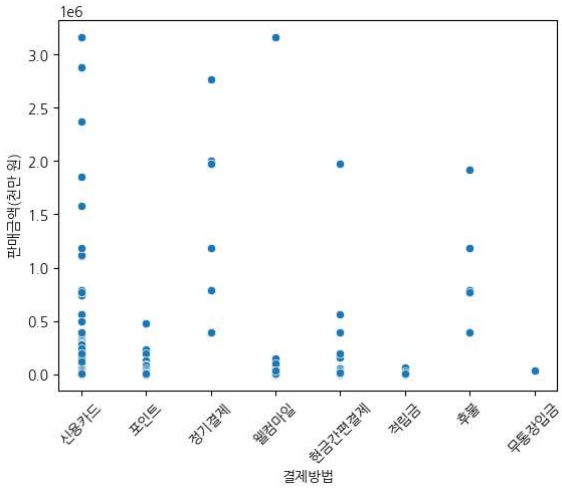
10월 매출이 가장 큰 폭으로 낮아진 것으로 보아 10월에 가장 많은 할부 결제가 있었음을 알 수 있다.



[그림 7. 월별 순수익(처리 상황, 할부 기간 고려)]

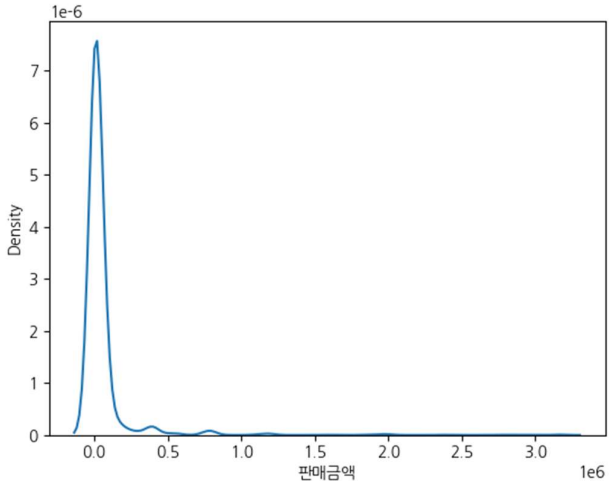
3. 결제방법에 따른 분석: 결제방법과 결제금액의 연관성 분석

Scatterplot을 통해 신용 거래는 점의 분포가 넓지만 현금 거래는 하단에 몰려 있는 양상을 통해 고액은 신용 거래를 통해서만 이루어졌음을 확인할 수 있다.



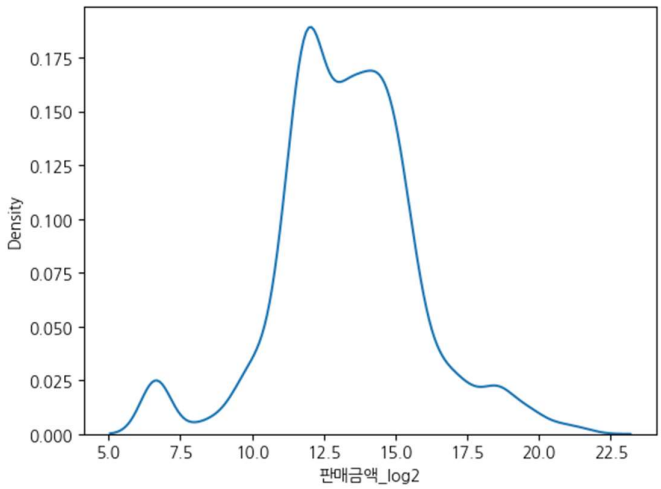
[그림 8. 결제방법에 따른 분석: 결제방법과 결제금액의 연관성 분석]

판매금액 분포를 정규화 하기 위해 판매금액을 로그 스케일로 변환하였다. 그 결과, 정규 분포 형태가 아니므로 shapiro-wilk 정규성 검정을 진행할 수 없다.



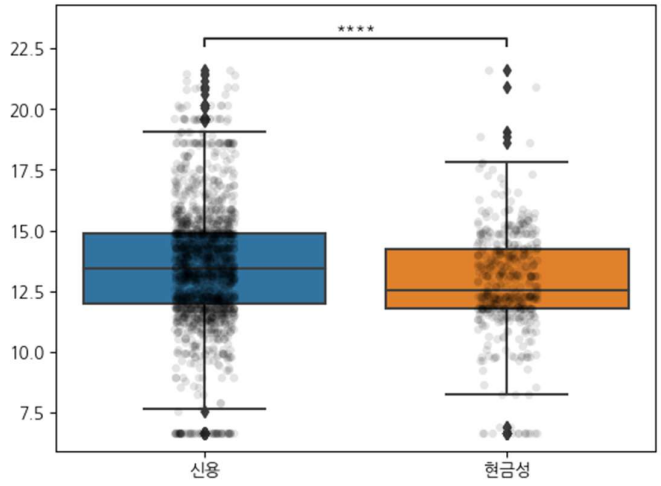
[그림 9. 판매금액에 따른 밀도 그래프]

1% 표본에 대한 밀도 그래프 및 정규성 검정을 진행 했을 때도 그래프는 정규분포로 보이지 않고, shapiro-wilk 정규성 검정을 진행했을 때 p-value가 0.05 이하로 나와 정규분포를 만족하지 않음을 확인하였다.



[그림 10. 1% 표본에 대한 밀도 그래프 및 정규성 검정 결과]

1% 표본의 결제방법을 신용 거래(신용카드, 후불, 정기 결제)와 현금 거래(그 외)로 나누어 두 그룹 간 판매금액의 차이를 확인하였다. 두 그룹의 건 당 판매금액으로 box plot과 scatter plot을 그리고 mann-whitney test를 진행하였고, P-value는 9.121e-07로 나타났다.



[그림 11. 신용 거래와 현금 거래 간의 판매금액 차이에 대한 box plot 및 scatter plot]

IV 토의

1. 결과 요약

가. 입점 기업별 RFM 분석(3그룹)

총 매출액이 가장 큰 상위 세 그룹인 다우기술, 지니, 천재 태블릿에 대한 RFM 분석을 진행하였다. R, F, M 분석 결과 모두 다우기술이 50% 넘는 큰 값을 보였고 지니와 천재 태블릿이 그 뒤를 이었다. 이를 통해 다우기술, 지니, 천재 태블릿 순으로 최근에, 자주, 많이 판매하였음을 확인하였다.

이 결과를 바탕으로 세 그룹을 기준으로 하면 충성 고객인 다우기술에 큰 혜택을 부여하고, 지니에는 혜택과 동시에 촉진 전략을 수행하고, 큰 차이로 3순위에 위치했던 천재 태블릿에는 큰 촉진 전략을 진행할 수 있다.

나. 매출 시각화

1) 월별, 연도별 매출

월별로는 3월이 가장 큰 매출을 기록하였고, 다음으로 5월과 10월이 높은 매출을 보였다. 연도별로는 2019년부터 2022년까지 꾸준한 매출 상승 추세를 보였다.

매출이 급증하는 3, 5, 10월에는 많이 판매되는 상품의 프로모션을 진행하여 고객들이 다른 상품과 함께 구매할 수 있도록 유도하고, 매출이 높지 않은 월에는 꾸준히 판매되는 상품에 대한 프로모션을 진행할 수 있다. 연도별로는 매출 상승 추세를 보이고 있으나, 추후 성장이 지체될 수 있음을 고려하여 마케팅 전략을 세워야 한다.

2) 월별 순수익 (처리 상황, 할부기간 고려)

처리 상황이 '구매 확정'된 행만을 대상으로 일시불, 1개월, 6개월, 12개월, 18개월, 24개월의 할부 기간을 계산하여 월별 순수익을 계산하였다. 그 결과, 10월의 할부 결제가 가장 많아 매출이 큰 폭으로 낮아졌음을 확인하였다. 이는 다음 해를 준비하며 금액이 높은 물품을 구매하면서 할부를 많이 진행했음을 유추해 볼 수 있다.

다. 결제방법에 따른 분석: 결제방법과 결제금액의 연관성 분석

신용 거래와 현금 거래의 결제금액 차이를 확인하기 위해 Scatterplot으로 시각화하였다. 신용 거래는 매출 분포가 넓게 퍼져 있는 반면, 현금 거래는 낮은 금액에 몰

려 있는 양상을 보였다.

데이터 정규화를 위해 판매금액을 로그 스케일 변환했을 때 정규분포 형태가 아님을 확인하였다. 정규분포가 아니기 때문에 Mann-Whitney U Test를 진행하였고 P-value가 9.121e-07라는 작은 값으로 나타나 신용 거래와 현금 거래 간 판매금액 차이는 통계적으로 유의미함을 확인하였다.

2. 한계점

본 분석의 한계는 다음과 같다.

먼저, RFM 분석 자체의 한계가 있다. RFM 분석은 다양한 시사점을 제공하지만 R, F, M을 기준으로 분류하는 것 외에는 정해진 것이 없기 때문에 실제 서비스에 적용하기 위해서는 추가 분석이 필요하다.

다음으로 데이터로 인한 한계가 있다. '할부기간'을 일시불, 1개월, 6개월, 12개월, 18개월, 24개월로 구분하여 분석을 진행하였으나 1개월 할부는 일시불로 취급하는 경우도 있기 때문에 분류에서의 모호함이 있었다. 1개월을 일시불과 같다고 보고 분석을 진행하면 결과가 달라지므로 이 부분을 명확히 할 필요가 있다.

마지막으로, 선택 분석 미실시로 인한 한계가 있다. 선택 분석을 진행하면 최대 매출 상품 세 종류 집계, 주문 연도에 따른 해당 상품의 매출 증감 분석, 주문한 달과 판매금액의 상관관계 분석을 진행하였을 것이다. 그러나 시간 부족으로 선택 분석을 진행하지 못했고, 필수 분석 결과를 기반으로 시사점을 도출할 때 한계가 있었다.

3. 제언

본 분석은 천재교육의 '프로젝트 기반 빅데이터 서비스 개발자 양성 과정'에서 본사로부터 제공 받은 '미니프로젝트-쇼핑몰 실습데이터.xlsx' 데이터를 활용하였다. 이 보고서는 다음의 경우 전략 수립의 근거로 활용 가능할 것으로 기대한다.

먼저, 입점 기업별 RFM 분석을 토대로 전략을 수립할 수 있다. 다음으로 매출 시각화 결과를 통한 전략을 수립할 수 있다. 마지막으로 신용 거래와 현금 거래 간 판매금액 차이를 바탕으로 신용 거래 활성화 전략과 현금 거래 유도 전략을 필요에 맞게 진행할 수 있다.

V 참고문헌

박광호. (2002). 인터넷 소매유통업의 RFM 모델 기반 충성고객관리를 위한 웹서비스 (WsLCM) 프레임웍 (Web services Framework for Loyal Customer Management based on RFM Models in Internet Retailing). 지능정보연구, 8(1), 41-63.

지현정, 신경일, 신동일, 신동규. (2017). RFM 기법과 K-Means 알고리즘을 이용한 고객 분류. 한국정보처리학회 학술대회논문집, 24(2), 803-806.

코드: https://github.com/LeeShinYoung00/my-git/tree/main