**쇼핑몰 데이터 분석 보고서**

1(팀)

이름: 김지석, 박지석, 허현강

1. **요약**

온라인 쇼핑몰의 매출을 증가시키기 위해 쇼핑몰의 데이터를 세 가지 관점에서 분석했다. 첫 번째로는 입점 기업별 RFM 분석을 통해 알파, 지니, 천재태블릿, 쥬크박스 등의 14개사 업체가 고객 가치가 높음을 확인했다. 두 번째로는 매출 시각화를 통해 특정 월에 매출이 증가하는 경향이 나타났다. 또한, 전체적으로 매년 매출이 증가하며 쇼핑몰이 성장하는 모습 나타냈다. 마지막으로 결제 방법에 따른 판매금액과의 연관성을 확인해보았을 때, 연관성이 없음을 확인했다.

위 세 가지 관점 이외에 매출 증가의 원인을 찾기 위해 추가로 코로나 확진자 수와 매출간의 관계성을 분석했고, 연관성이 있음을 확인했다.

1. **서론**
2. 프로젝트의 목적
3. 쇼핑몰의 매출을 증가시키기 위해 세 가지 관점으로 분석한다.
4. 입점 기업별 RFM 분석: 고객을 최근 구매일(Recency), 구매 빈도(Frequency), 구매 금액(Monetary)의 기준으로 점수를 매겨 가치가 높은 그룹, 가치가 중간인 그룹, 가치가 낮은 그룹으로 분류하고, 각 그룹마다 매출을 증가시킬 수 있는 개별적인 판매전략을 세운다.
5. 매출 시각화: 월별 및 연도별 매출 데이터를 시각화 하여 시간의 흐름에 따른 매출 추이를 분석한다. 이를 통해 매출과 순수익 증대시기를 파악하여 쇼핑몰의 마케팅 전략을 세운다.
6. 결제방법과 결제 금액간 연관성 분석: 결제 방법과 결제 금액 간의 연관성을 분석해 어떤 결제 방법이 더 높은 매출로 나타나는지 파악하여 해당 결제 방법을 더 사용하게 만들 수 있는 전략을 세운다.
7. 추가분석
8. 코로나 확진자와 쇼핑몰 매출간의 상관관계 분석: 월별로 코로나 확진자의 증감과 쇼핑몰 매출 증감간의 상관관계를 분석하여 이에 대응하는 전략을 세운다.
9. 가설
10. [RFM 분석 가설]
11. 최근에 구매한 고객이 더 자주, 더 많은 금액을 구매할 것이다.
12. 구매 빈도가 높은 고객이 더 많은 매출을 창출할 것이다.
13. 높은 금액을 지출한 고객이 전체 매출에 더 큰 기여를 할 것이다.
14. R, F, M 점수가 높은 고객은 매출에 기여한 바가 높을 것이다.
15. [매출 시각화 가설]
16. 특정 시기, 월, 연도에 따라 매출이 다르게 나타날 것이다.
17. [결제 방법과 매출 간의 연관성 가설]
18. 특정 결제 방법을 사용하는 고객이 더 높은 금액을 결제할 것이다.
19. [매출 금액과 코로나19 간의 연관성 가설]
20. 코로나19와 매출 간의 관계가 있을 것이다.
21. 검증된 내용
22. 중심극한정리

표본의 크기가 30 정도가 되면, 표본 평균들의 분포는 표준정규분포에 가까워지게 된다. 여기에서, 모집단의 분포에 대한 가정은 무의미해진다. 중심극한정리에 따르면 표본의 크기가 큰 경우 모집단의 분포에 관계없이 표본평균들의 분포가 정규분포를 이루기 때문이다. (Kwak et al., 2017)

1. **분석 방법**
2. 순서도

|  |
| --- |
| [그림 1] 순서도 |

1. 데이터 수집
   1. 천재교육 쇼핑몰의 2019년 12월 13일부터 2022년 11월 8일까지의 데이터를 기반으로 진행하였으며, 프로젝트에 사용된 데이터는 다음과 같다.
      * 1. '주문번호','판매금액','결제방법','주문일자','처리상태','할부기간 데이터를 수집
   2. 감염병포털에서 2019년 8월1일부터 2023년 8월 31일까지의 코로나19 데이터를 수집하였으며, 프로젝트에 사용된 데이터는 다음과 같다.
      * 1. '일자', '국내발생(명)' 데이터를 수집
2. 데이터 전처리
   1. 데이터 클리닝
      * 1. 관련 없는 데이터 삭제(셀러리스트, 초도상품, 제작문구내역, 제조사, 판매금액이 0원인 제품)
   2. 데이터 변환
      * 1. 분리저장(할부 데이터 월별로 분할), 단위변환, 형태변환
   3. 데이터 통합
      * 1. 동일 의미의 결제방법, 처리상태 항목 통합 처리
        2. 매출의 기준을 고객이 쇼핑몰에서 일단 구매를 했다면, 금액을 결제한 것으로 보고 주문취소, 환불완료 등을 제외한 나머지를 매출로 집계함. (구매확정, 교환완료, 배송준비, 배송중, 배송완료, 부분취소, 부분환불, 환불요청)
        3. 순이익은 구매확정된 제품만 집계함.
3. 사용한 분석 도구
   1. RFM분석

Recency('주문일자'): 최근 구매일로부터 경과한 일수.

값이 낮을수록 최근에 구매한 고객을 의미한다.

R 값이 낮은 고객은 최근에 활동한 고객으로 간주된다.

Frequency('주문번호'): 특정 기간 동안의 구매 횟수.

값이 높을수록 구매 빈도가 높은 고객을 의미한다.

F 값이 높은 고객은 자주 구매하는 충성 고객으로 간주된다.

Monetary('판매금액'): 특정 기간 동안의 총 구매 금액.

값이 높을수록 지출이 많은 고객을 의미한다.

M 값이 높은 고객은 높은 금액을 지출하는 가치 높은 고객으로 간주된다.

분류기준

높은 그룹 (9점): R, F, M 점수의 합이 9점인 고객.

중간 그룹 (6-8점): R, F, M 점수의 합이 6-8점인 고객.

저가 그룹 (1-5점): R, F, M 점수의 합이 1-5점인 고객.

* 1. 카이제곱 (독립성 검정)

두 범주형 변수 사이에 독립적인 관계가 있는지를 검정하는 분석

귀무가설(H0): 두 변수는 독립적이다.

대립가설(H1): 두 변수는 독립적이지 않다.

* 1. 표본의 수 확인

모집단의 분포가 어떤 분포라도 추출된 표본의 수가 충분히 크다면 표본평균의 분포는 정규분포에 근사한다. 따라서 판매금액은 정규성을 띈다고 가정한다.

* 1. T-test

각 결제 방법 쌍에 대해 독립 표본 T-검정을 수행하여 t-통계량과 p-값을 계산 한다.

p-값이 작을수록 두 집단 간의 평균 차이가 통계적으로 유의미함을 나타낸다.

* 1. 피어슨 분석

두 연속형 변수 간의 선형 관계를 측정하는 통계적 방법으로 피어슨 상관계수는 -1에서 1 사이의 값을 가지며, 두 변수 간의 관계가 얼마나 강한지를 알 수 있다.

0.7 ≤ |r| ≤ 1.0: 강한 상관관계

0.3 ≤ |r| < 0.7: 중간 정도의 상관관계

0.0 ≤ |r| < 0.3: 약한 상관관계

피어슨 분석의 가정은 다음과 같다.

선형성: 두 변수 간의 관계가 선형적이어야 합니다.

정규성: 각 변수는 정규 분포를 따라야 합니다.

동분산성(Homoscedasticity): 각 변수의 분산이 일정해야 합니다.

독립성: 각 관측값은 독립적이어야 합니다.

1. **결론**
2. 입점 기업별 RFM 분석

|  |
| --- |
| [표 1] RFM분석표 |
| [그림 2] RFM 분석그래프 - Recency Histogram    X축: Recency (최근 구매일로부터 경과한 일수)  Y축: Count (고객 수)  1) Recency Histogram의 경우 분포가 비대칭으로 왼쪽에 치우쳐져 있다. 이는 많은 고객이 최근에 구매했음을 나타내며, 시간이 지남에 따라 구매한 고객의 수가 급격하게 줄어든다는 것을 보여준다.  [그림 3] RFM 분석그래프 - Frequency Histogram    X축: Frequency (구매 횟수)  Y축: Count (고객 수)  2) Frequency Histogram그래프의 경우 대부분의 고객은 구매 횟수가 매우 적고, 소수의 고객만이 빈번하게 구매하며, 20,000회 이상의 구매 기록을 가진 고객은 거의 없다.  분포는 매우 비대칭적으로, 대부분의 고객이 낮은 구매 빈도를 가지고 있다.  [그림 4] RFM 분석그래프 - Monetary Histogram    X축: Monetary (총 구매 금액)  Y축: Count (고객 수)  3) Monetary Histogram의 경우 대부분의 고객이 소비한 금액은 매우 낮다.  소수의 고객만이 매우 높은 금액을 소비했다. 또한 분포는 역시 비대칭적으로, 대부분의 고객이 낮은 구매 금액을 기록하고 있다.  4)위 RFM분석 그래프들을 종합적으로 해석해보면  많은 고객이 최근에 구매를 한 것으로 보아 고객 유지가 잘 되고 있음을 알 수 있다. 대부분의 고객이 낮은 빈도로 구매하고 적은 금액을 소비하고 있다. 이는 많은 고객이 일회성 구매자이거나 가끔 소액으로 구매하는 고객임을 알 수 있다. 일부 고객만이 높은 금액을 소비하고 있어 이 고객들이 전체 매출에 큰 영향을 미치고 있음을 나타낸다.  대표적으로 알파, 지니, 천재태블릿, 쥬크박스 등의 14개 업체가 고객가치가 높음을 알 수 있다. |

1. 월별, 연도별 매출 시각화

|  |
| --- |
| [그림 5] 월별 매출 그래프    x축: 월(2019년 12월 ~ 2022년 11월)  y축: 매출(백만 원 단위)   1. 2019년부터 2020년 9월까지는 매출이 낮은 상태를 유지하다가 2020년 10월에 400.7백만원으로 폭발적으로 매출이 증가하였다. 이후로 점진적으로 우상향 그래프를 그리다가, 2022년 3월과 5월에 각각 723.4백만원, 838.0백만원으로 최고 매출을 기록하였다. 이후로는 우하향 방향으로 나타나고 있다. |
| [그림 6] 연도별 매출 그래프    x축: 연도(2019년 ~ 2022년)  y축: 매출(백만 원 단위)   1. 2019년의 경우 12월 한 달의 매출만 집계되어 매우 적은 매출을 나타낸다. 2020년부터는 매년 최소 1.5배 이상 매출이 증가하고 있음을 알 수 있다. |

1. 월별, 연도별 순수익 시각화

|  |
| --- |
| [그림 7] 월별 순수익 그래프    X축: 연도-월 (2019년 12월 ~ 2024년 9월까지)  Y축: 매출 (백만 원 단위)   1. 월별 매출그래프와 비슷하게 2019년 12월부터 2021년 1월까지 저조한 매출을 보이다가, 2021년 2월에 609.1백만원으로 순이익이 크게 증가함을 알 수 있다. 이후로는 매출이 여러 차례 급증하며, 2022년 3월에 917.9백만원, 2022년 5월에 1014.5백만원으로 최고 순이익을 기록하였다. 하지만, 이후로는 매출이 꾸준히 감소하는 경향을 보였다.   2022년 11월 이후의 순이익은 할부데이터가 상계된 값으로, 실제 데이터와는 상이하다. |
| [그림 8] 연도별 순수익 그래프    X축: 연도 (2019년부터 2024년까지)  Y축: 매출 (백만 원 단위)   1. 연도별 매출 그래프와 마찬가지로 19년도에는 12월밖에 집계되지 않아 매출이 적게 집계되었다. 2020년부터는 최소 1.6배 이상의 증가폭을 보이며 매출이 증가하였다. 2022년 11월 이후의 순이익은 할부데이터가 상계된 값으로, 실제 데이터와는 상이하다. |

1. 결제 방법과 결제 금액의 연관성 분석

\* 결제 방법은 맨 앞의 한 가지만 사용한 것으로 간주

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [그림 9] 결제 방법별 매출 그래프   1. 결제 방법별 매출 그래프     X축: 결제방법  Y축: 매출 (백만 원 단위)  [표 2] 결제 방법별 매출표   |  |  | | --- | --- | | **결제방법** | **판매금액(원)** | | 기타 | 1,127,845,130 | | 마일리지 | 776,713,165 | | 신용카드 | 7,229,664,135 | | 할부 | 1,981,372,200 | | 현금 | 408,470,005 |  1. 신용카드를 이용한 결제 방법이 약 72억으로 가장 높은 결제금액을 보여주었고, 할부 – 기타 – 마일리지 – 현금 순으로 결제 금액이 점점 낮아졌다.   [표 3] 결제 방법별 표본수   |  |  | | --- | --- | | **결제 방법** | **표본수(개)** | | 기타 | 1,361 | | 마일리지 | 41,839 | | 신용카드 | 166,021 | | 할부 | 1,938 | | 현금 | 7,396 |  1. 중심극한정리   중심극한정리에 따르면 표본의 크기가 큰 경우 모집단의 분포에 관계없이 표본평균들의 분포가 정규분포를 이룬다. 결제 수단 별 판매금액이 집계된 데이터의 수량을 보면 최소 1300개 이상이므로, 표본의 크기가 충분히 커 표본평균의 분포가 정규분포를 이룬다고 가정한다  [표 4] T-test   |  |  |  | | --- | --- | --- | | **결제 수단별 비교** | **T-test: t-statistic** | **P-value** | | 현금 vs 신용카드 | 4.314955239752251 | 1.616085814964818e-05 | | 현금 vs 할부 | -41.42019064057588 | 6.395592277807957e-271 | | 현금 vs 마일리지 | 13.631770363564108 | 8.04088780354323e-42 | | 현금 vs 기타 | -62.85908819432712 | 0.0 | | 할부 vs 마일리지 | 43.27116596415076 | 8.168294374647739e-287 | | 할부 vs 신용카드 | 42.19052061471383 | 1.6464308003796388e-276 | | 마일리지 vs 신용카드 | -47.44191806644744 | 0.0 | | 마일리지 vs 기타 | -67.42782886211576 | 0.0 | | 기타 vs 신용카드 | 65.32693106243991 | 0.0 | | 기타 vs 할부 | -7.4150310693776005 | 1.5986315011814252e-13 |  1. 모든 비교에서 p-값이 매우 작아 결제 방식 간의 평균 판매금액 차이가 통계적으로 유의미한 것으로 나타났다. 이는 각 결제 방식이 판매 금액에 미치는 영향이 다르며, 이를 고려한 마케팅 전략이 필요하다는 것을 의미한다. |
| [그림 10] 신용카드 결제와 매출의 상관관계 - 피어슨분석    X축: 결제방법-신용카드  Y축: 매출  [그림 11] 현금 결제와 매출의 상관관계 - 피어슨분석    X축: 결제방법-신용카드  Y축: 매출  [그림 12] 할부 결제와 매출의 상관관계 - 피어슨분석    X축: 결제방법-할부  Y축: 매출  [그림 13] 마일리지 결제와 매출의 상관관계 - 피어슨분석    X축: 결제방법-마일리지  Y축: 매출  [그림 14] 기타 결제와 매출의 상관관계 - 피어슨분석    X축: 결제방법-기타  Y축: 매출   1. 각 결제방법 별 피어슨분석을 진행한 결과, P값은 모두 낮게 나와 유의미한 분석임을 확인했다. 하지만 모두 R값이 0.1미만으로 나타나 낮은 연관성을 보였으며, 시각화 역시 유의미한 연관성이 보이지 않았다. |

1. 결과 해석

중심극한정리에 의해 표본평균이 정규성을 띈다고 가정한다.

이에 따라 T-test 결과, P-value 값이 0.05 미만으로 나타나 결제 방법에 따른 매출 금액의 차이가 유의미하다는 것을 확인했다.

추가로 피어슨 상관분석을 통해 상관관계를 분석한 결과, P-value값은 0.05 미만이였지만, 상관계수가 0.1 미만으로 사실상 관련성이 없다는 결론을 도출했다.

1. 월/연별 매출 증가에 영향을 미친 외부요소

|  |
| --- |
| [그림 15] 코로나 확진자 수와 매출 금액 비교 그래프    X축: 연도-월 (2021년 8월 ~ 2022년 12월까지)  y축(좌): 확진자 수 (백만 명 단위)  x축(우): 매출금액 (억 원 단위)   1. 2022-1월부터 2022-9월까지 확진자 수와 매출이 비슷한 증감을 보인다. 이는 코로나 확진자 수와 매출 금액이 연관성이 있음을 예측 할 수 있다. |
| [그림 11] 코로나 확진자 수와 매출 금액 상관관계 그래프(피어슨 분석)    X축: 코로나 확진자수(? 명 단위)  y축: 판매금액 (백만 명 단위)   1. 이를 상관관계 분석법으로 확인해보았을 때, 상관계수가 0.81, p-value가 3.89e-09으로 높은 상관 관계를 확인하였다. 2. 결과 해석   내부 연관성 분석에서 유의미한 결과를 얻지 못했기 때문에, 매출의 증감과 연관된 외부 요인에 대한 탐색을 진행했다. 분석 시기가 코로나19 시기와 겹치는 것을 확인하고, 코로나 확진 동향과 매출 간의 상관관계를 분석했다. 그 결과, 로그 변환된 코로나 확진자 수와 매출 금액 간의 상관관계에서 P-value가 0.81로 큰 유의미한 결과를 도출했다. |

1. **토론**
2. RFM 분석 결과를 확인했을 때, 쇼핑몰의 매출을 증가시키기 위해서는 RFM 점수가 높은 고객들에게 적당한 이벤트를 통해 충성고객으로 유지시키는 것이 중요하고, 점수가 중간인 고객들과 낮은 고객들에게 다시 주문을 받기 위해 웰컴백 쿠폰, 적립금 할인 등의 적극적인 마케팅을 활용해 유치한다면 매출 증가에 도움이 될 것이다.
3. 특정 시즌별로 매출이 증가하기 때문에 해당 시즌에는 기획전을 통해 타 업체보다 조금 더 싼 가격으로 고객들을 유치한다거나, 구매 인증 이벤트를 진행하는 등의 시도를 통해 매출을 증가시킬 수 있을 것이다.
4. 현재 결제 방법과 매출액의 연관성은 없지만, 신용카드로 결제했을 때 가장 많은 매출액을 기록했으므로, 카드사와 제휴를 통해 일정 금액 이상 신용카드 할인 이벤트 등을 진행해 신용카드로 결제를 유도한다면 매출 증가에 도움이 될 것이다.
5. 그리고, 앞으로 비대면 상황이 다시 발생한다면, 발빠르게 비대면 맞춤 상품들을 추천 배너에 등록함으로써 고객들이 쉽게 접근할 수 있게 만든다면, 매출이 증가할 것이다
6. 완성도 높은 데이터 분석을 위해서는 외부 변수에 의향 영향도 함께 고려해야 함을 이번 프로젝트를 통해 배울 수 있었다.
7. 추가적인 분석 주제로는 최대 매출 상품을 집계해보고 이 상품들이 주문 연도에 따른 주문량 증감을 분석해볼 수 있을 것이다.
8. **참고문헌**

* 윤석모. (2021). 온라인 쇼핑몰 RFM(Recency, Frequency, Monetary)모형 기반 군집분석에 따른 연관성 규칙(Association Rule) 전략 연구 [석사학위논문, 충북대학교]. <http://www.riss.kr/link?id=T15916680>
* 질병관리청.(n.d.).코로나19양성자(표본)감시현황.

<https://dportal.kdca.go.kr/pot/cv/trend/dmstc/selectMntrgSttus.do>

* Kwak, S. G., J.H. Kim., “Central limit theorem: the cornerstone of modern statistics”, Korean journal of anesthesiology, Vol 70, No 2 (2017), 148

|  |  |
| --- | --- |
| **사용한 파이썬 및 패키지 버전** | |
| Python | #version – 3.11.7 |
| pandas | # version - 2.1.4 |
| seaborn | # version - 0.12.2 |
| numpy | # version - 1.26.4 |
| matplotlib | # version - 3.8.0 |
| ydata\_profiling | # version - 4.8.3 |
| statsmodels.api | # version - 0.14.0 |
| from openpyxl import Workbook | # version - 3.0.10 |
| from openpyxl.drawing.image import Image | # version - 3.0.10 |
| from openpyxl.utils.dataframe  import dataframe\_to\_rows | # version - 3.0.10 |
| from scipy import stats | # version - 1.11.4 |
| from statsmodels.formula.api import ols | # version - 0.14.0 |

