2025-07-17

**멋쟁이사자처럼**
FINAL PROJECT 5조

# 신용카드 이용 데이터 기반 고객 세그먼트 분석

실제 카드사 데이터를 활용한 고객 행동 특성과
세그먼트 등급 연관성 분석 및 모델 평가  프로젝트

조장 / 양유빈
조원 / 김태미
윤규남

# 목차

CONTENTS

# 01 탐색적 데이터 분석 (EDA)

## 분석 방향

고객 세그먼트(A~E) 특성 분석

### 분석 배경

고객 등급이 높을수록
활동성이 높을 것이라는
일반적 가정을
실제 데이터로 검증하고자 함

### 분석 방법

A~E 등급 고객을 대상으로
카드 사용 행태, 거래 빈도 등
주요 변수에 대한 가설 기반
EDA 수행

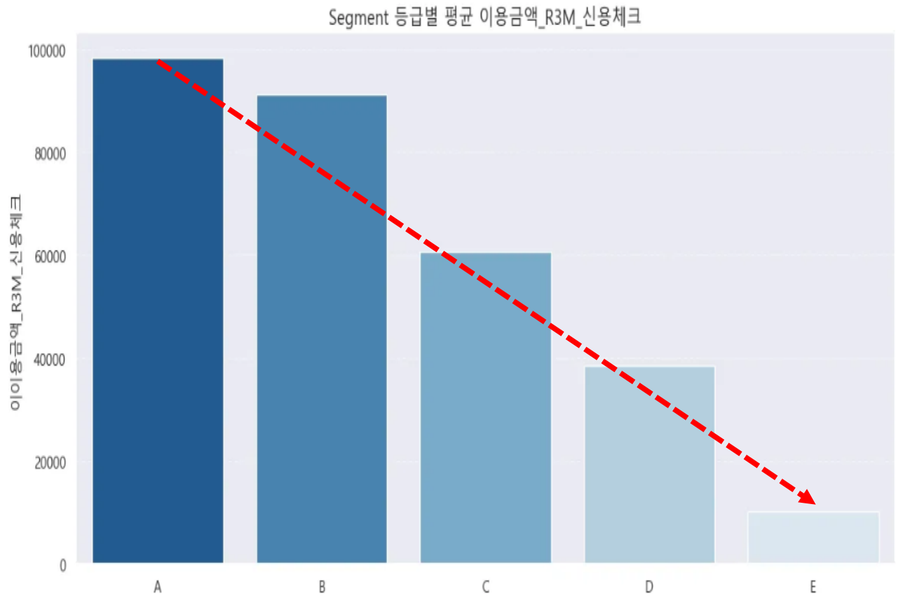
### 분석 목적

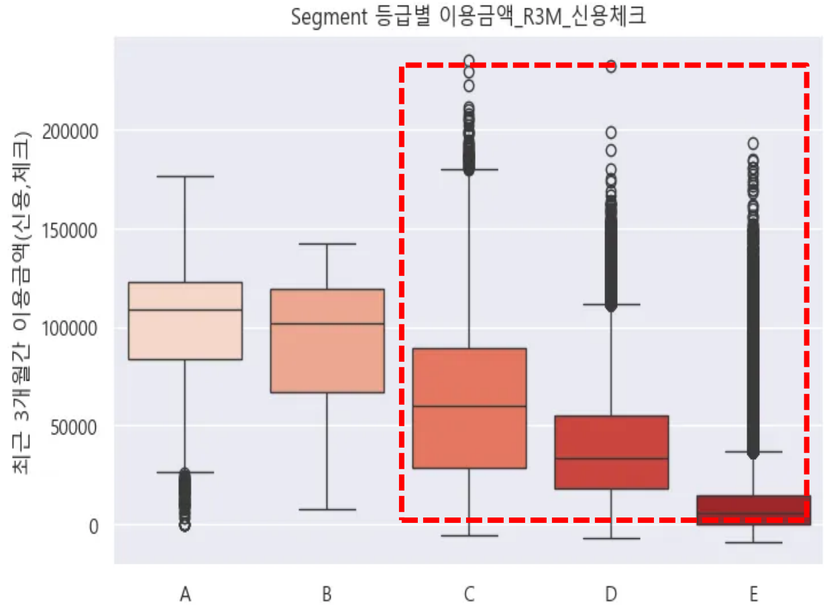
고객 등급 별 특성의 차이를
파악하여 등급 체계의
타당성과 세그먼트의 구조를
검토

# 01 탐색적 데이터 분석 (EDA)

## 가설 1.

최근 3개월간 신용/체크카드 이용금액이 높을수록 고객 등급이 높을 것이다



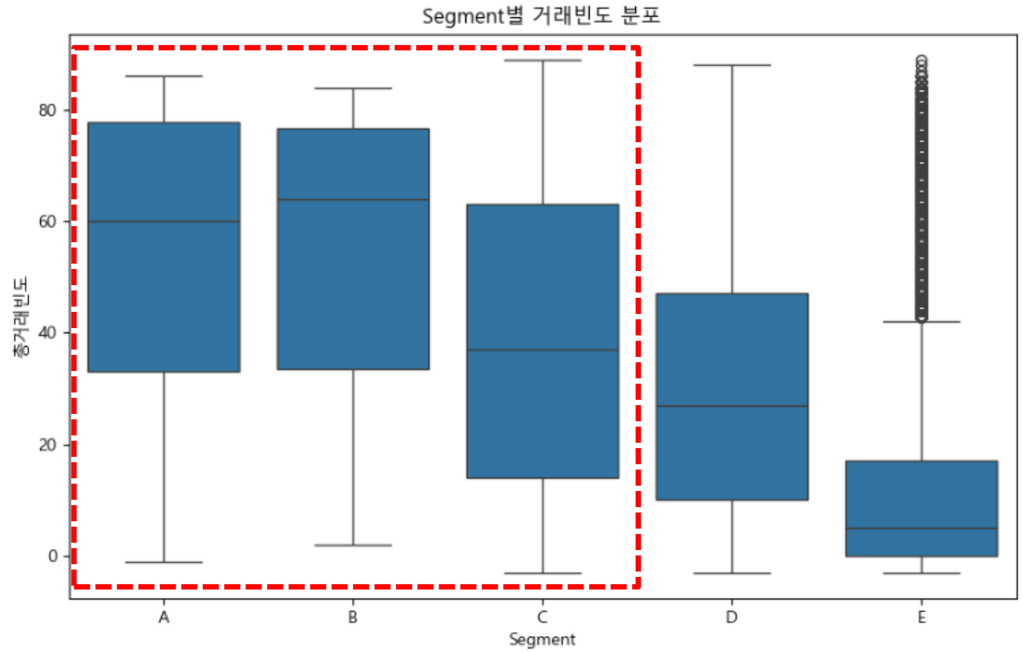




- 최근 3개월간 카드 이용금액은 **Segment 구분의 핵심 차별화 지표**로, 이용금액이 낮을수록 등급도 **일관되게 하락**함.
- A, B는 이용금액이 높은 고객, C, D는 중간수준이나 일부 높은 금액이 존재, E는 이용금액이 적은 고객이지만 데이터의 분포가 넓어 해석에 유의가 필요함.
- 다만 Segment A와 B의 그룹 간 차이는 사후검정 결과 통계적으로 유의하지 않아($p>0.05$) 해석에 주의가 필요함.

# 01 탐색적 데이터 분석 (EDA)

**가설 2.**

카드 사용 빈도가 증가할수록 고객 등급이 높아질 것이다.

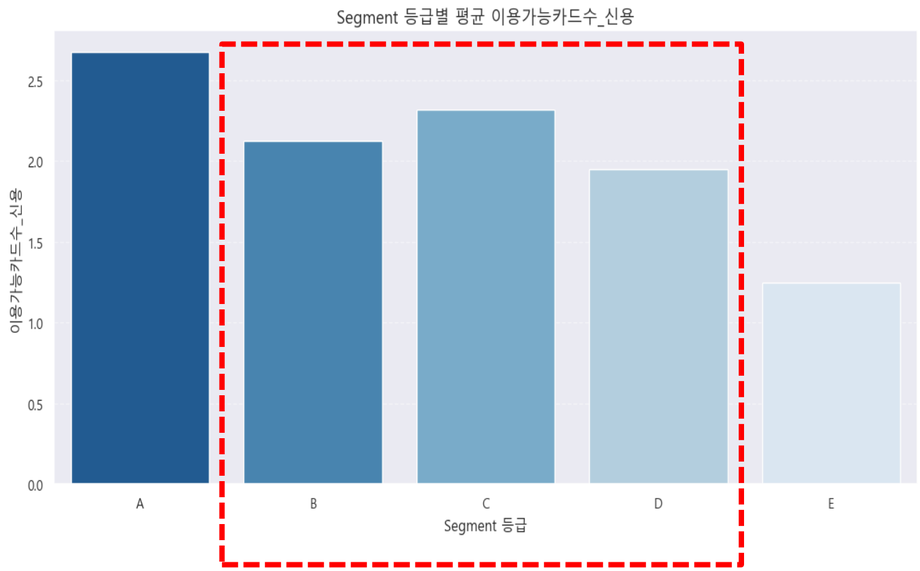


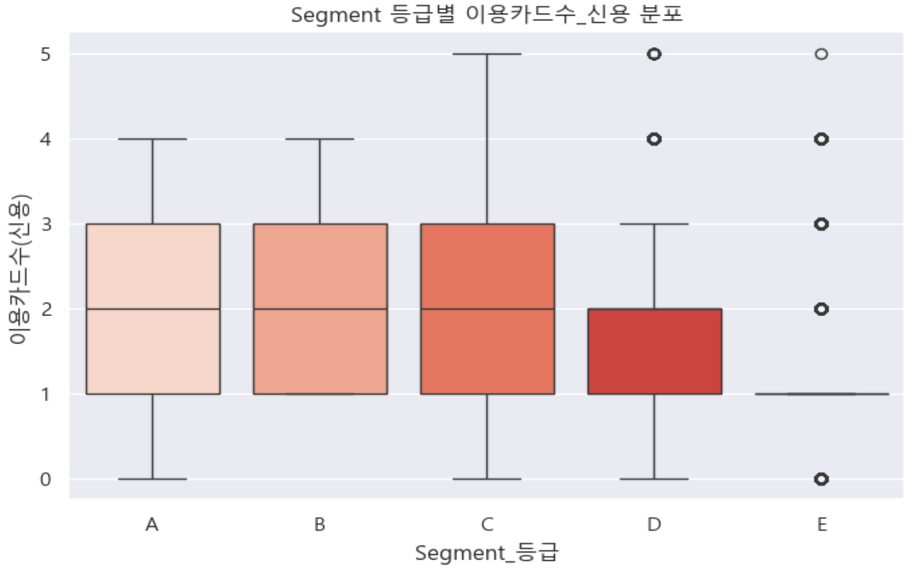

- A, B는 고빈도 거래 고객군, C는 중간 수준, D, E는 저빈도 거래군으로 거래빈도는 **고객 등급 간 순차적 차이**를 보임.
- 사후검정 결과 Segment A와 B, B와 C등급 간 카드 사용 빈도에 대한 차이가 유의하지 않아($p>0.05$) 해석에 주의가 필요함.

# 01 탐색적 데이터 분석 (EDA)

## 가설 3.

실제 사용 중인 신용카드 수가 많을수록 고객 등급이 높을 것이다



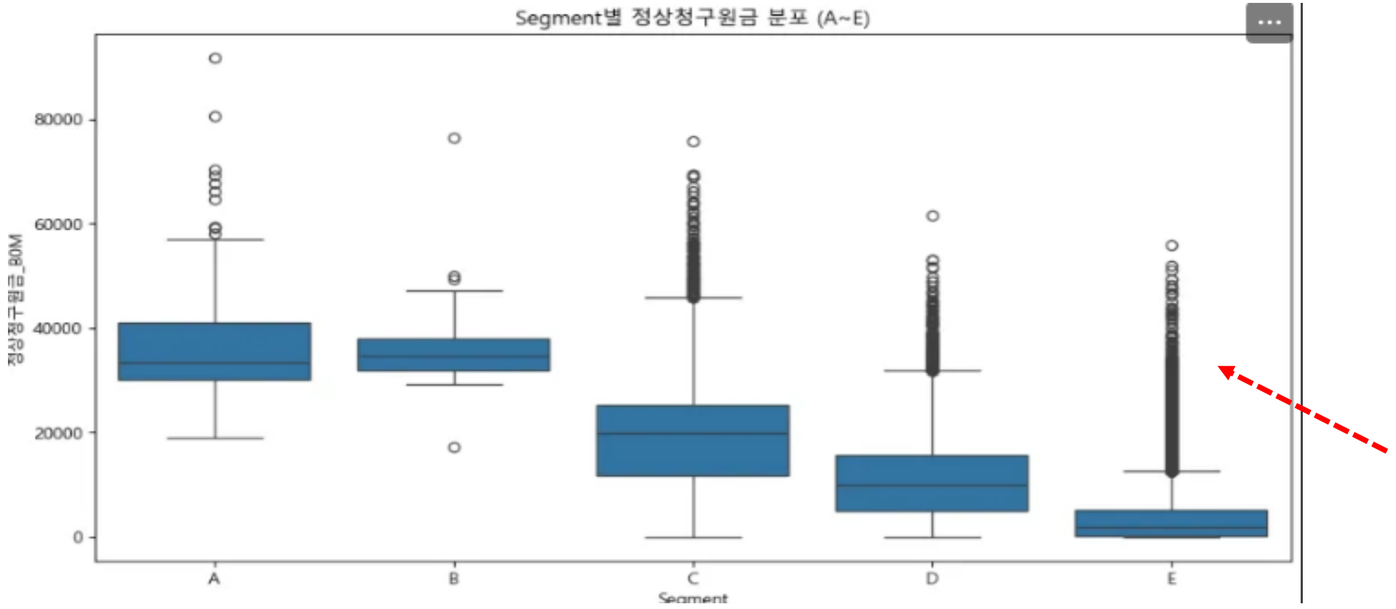




- Segment A에서 E로 갈수록 **실제 사용 중인 신용카드 수는 감소**하는 경향을 보임.
- 다만 Segment B와 C, B와 D의 그룹 간 차이는 사후검정 결과 통계적으로 유의하지 않아(p>0.05) 해석에 주의가 필요함.

# 01 탐색적 데이터 분석 (EDA)

**가설 4.**

월 청구금액이 높을수록 Segment 등급도 높을 것이다

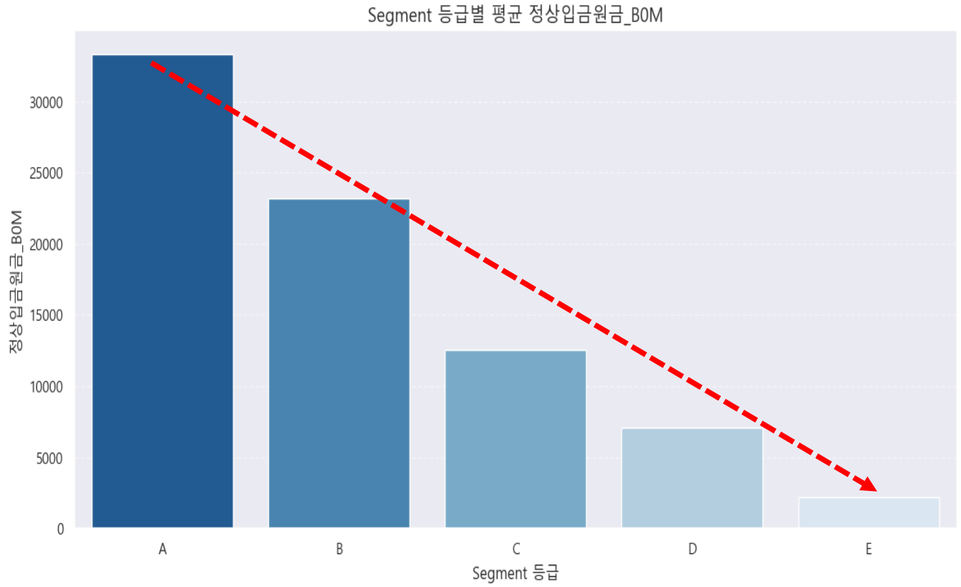


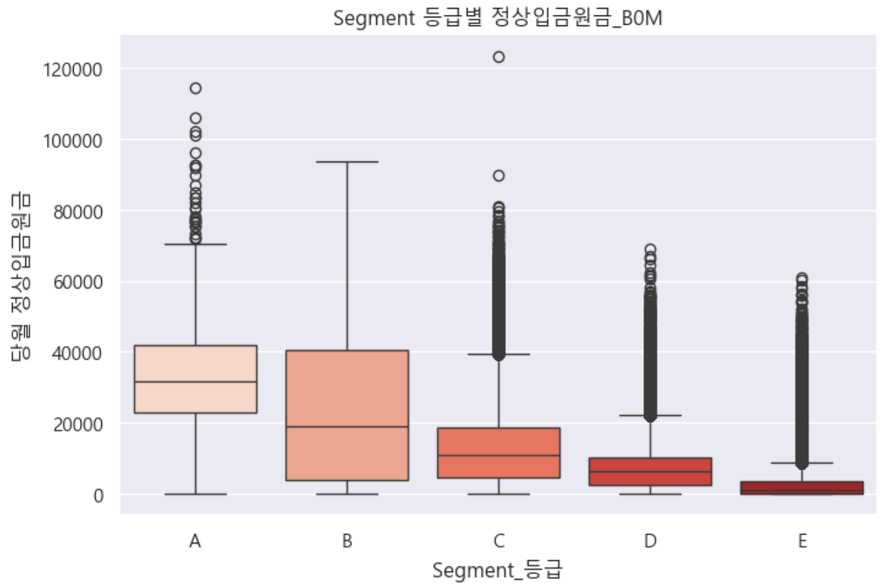

- 정상청구원금은 Segment 간 감소 추세를 보여 각 등급의 소비 규모 및 재정 건전성 차이를 반영함.
- 세그먼트 C, D, E에서는 **고액 극단치가 나타나** 일시적 이용 증가나 비정형적 소비 패턴의 가능성이 있음.
- Segment A와 B, B와 C간 '정상청구원금_B0M(당월 정상청구원금)'에 대한 차이가 통계적으로 유의하지 않음(p>0.05)

# 01 탐색적 데이터 분석 (EDA)

**가설 5.**

정상입금원금이 높을수록 고객이 상위 등급(Segment A 또는 B)에 속할 가능성이 높을 것이다



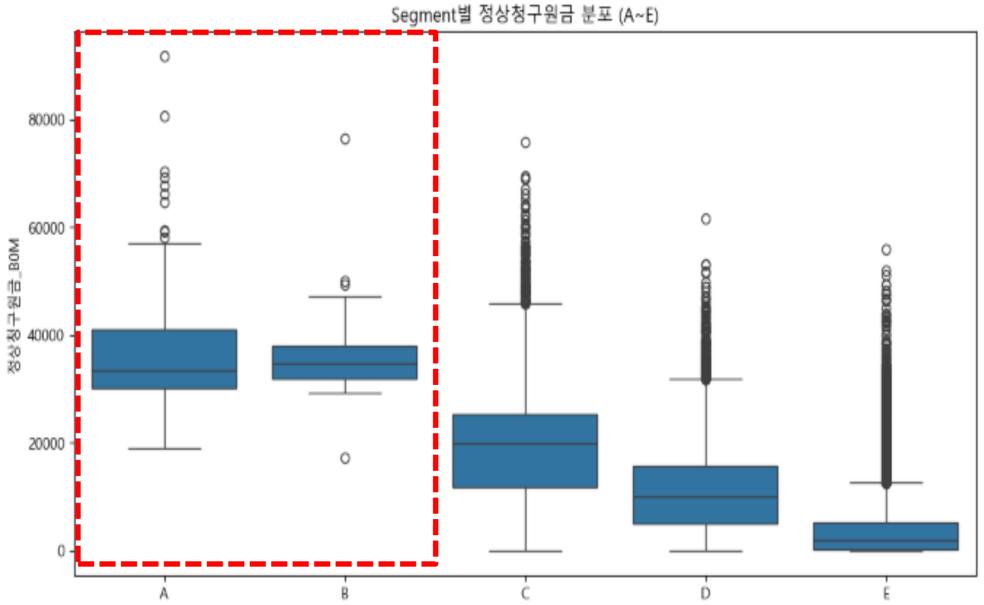



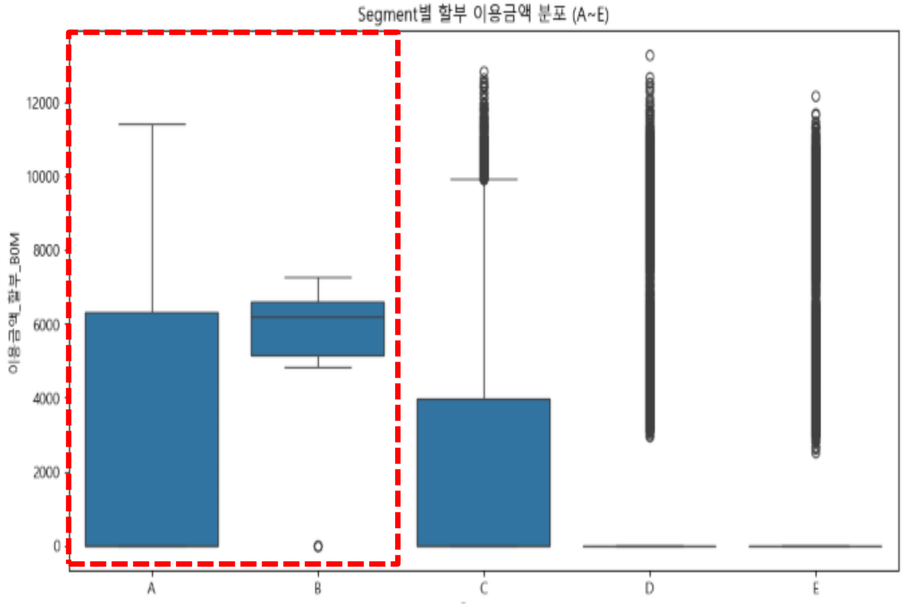

- 정상입금원금은 Segment A에서 E로 갈수록 **점진적으로 감소**하며, 전반적으로 상위 등급일수록 정상입금원금이 높은 경향을 보임.
- 다만 B와 C, B와 D등급 간에는 통계적으로 유의한 차이가 없어(p>0.05) 중간 등급 간 구분력은 제한적임.

# 01 탐색적 데이터 분석 (EDA)

**가설 6.**

카드사로부터 대출 금액이 클수록 고객 등급이 높을 것이다



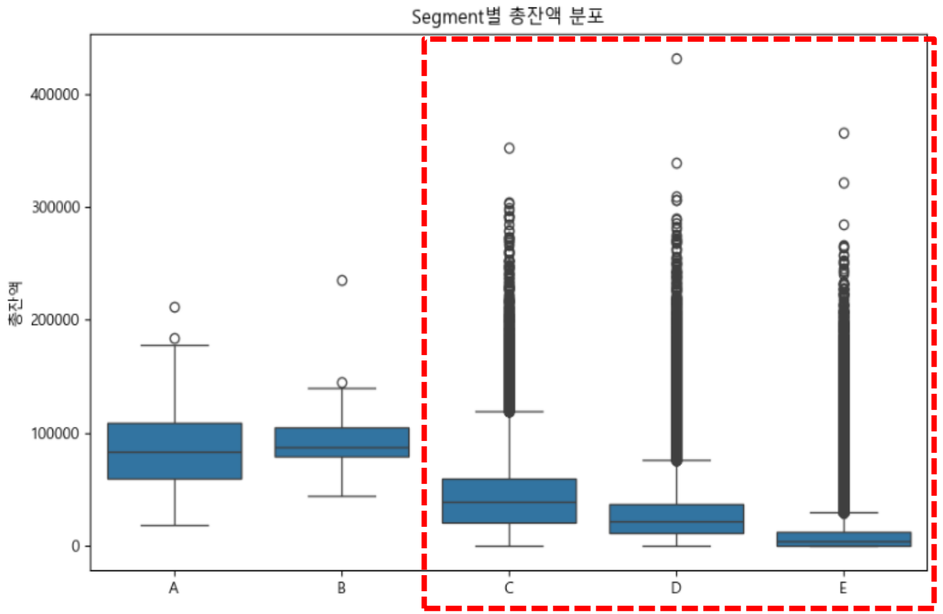



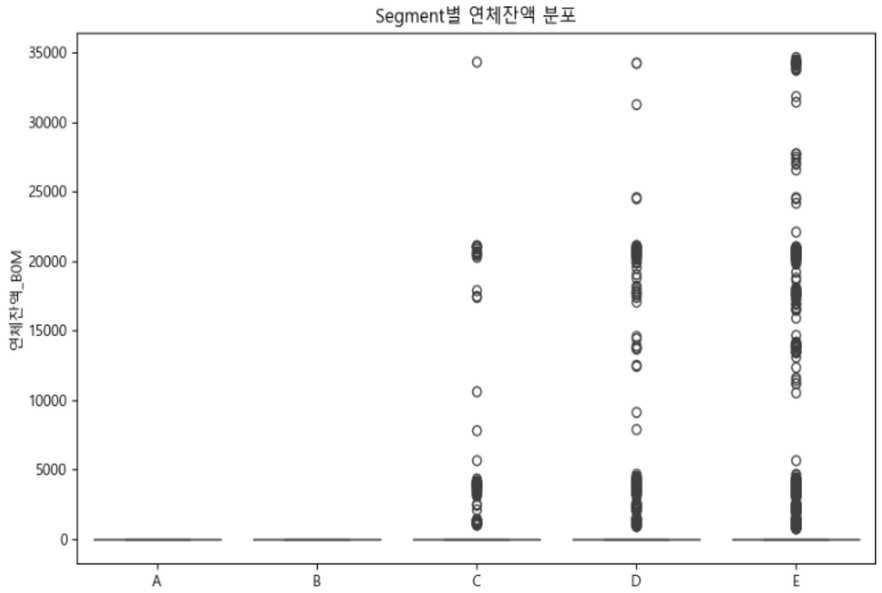

- Segment B는 A보다 정상청구원금과 할부 이용금액이 많아 소비성향이 더 강한 것을 확인할 수 있음.
- 모든 그룹들 간 '할부이용금액'에 대한 차이는 통계적으로 유의함($p<0.05$)
- 정상청구원금과 할부 이용금액 두 지표의 경우 A와 B를 구분할 수 있는 핵심 지표로 작용할 가능성이 있음 .

# 01 탐색적 데이터 분석 (EDA)

**가설 7.**

연체 잔액이 높은 고객은 주로 낮은 고객등급에 집중될 것이다




- Segment A와 B는 **높은 총 잔액임에도 불구하고 연체액이 거의 없어** 재정적으로 안정된 핵심 고객군으로 분류됨.
- 반면 Segment C~E는 총 잔액과 연체 잔액 모두에서 분포 편차가 크고 고액 연체자가 포함되어 있어 세그먼트 내에서 다양한 신용 위험 특성을 가진 고객군으로 해석할 수 있음 .
- 총 잔액에 대한 사후검정 결과, Segment A와 B 간의 차이는 통계적으로 유의하지 않은 것으로 나타남.

# 01 탐색적 데이터 분석 (EDA)

## 핵심 변수 도출 및 검증 절차

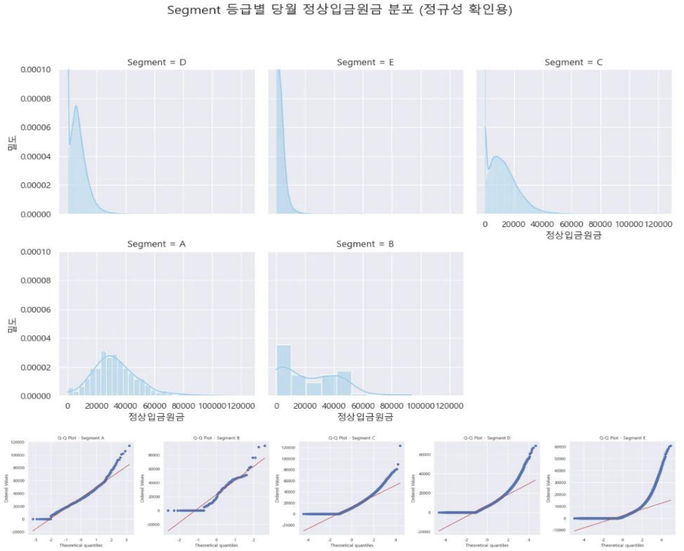
**1 단계**

- 고객 가치를 설명할 수 있는 **행동지표 기반 가설 변수**를 세그먼트별로 설정함.
- 변수 선정을 위해 Segment와 각 컬럼 간 **스피어만 상관분석**을 실시하고,
  |ρ| > 0.4인 변수를 유의미한 후보변수로 선정함.

**2 단계**

- Segment 간 차이를 검증하기 위해 **Kruskal-Wallis H-test, Dunn's test** 등을 수행함.
- 이 중 통계적으로 유의한 차이를 보인 변수를 머신러닝 모델에 포함할 핵심 피처로 선정함.
- 다만 통계적으로는 유의하지 않더라도 실무적 중요성이나 이론적 설득력이 있는 변수는 포함함.

(1) 정규성 확인




- 모든 등급들에 대하여 양의 왜도가 존재하며, 모든 등급들은 정규성을 만족하지 못한다.

(4) 사후검정(Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 2.323404e-15 | 1.486329e-79 | 7.380307e-177 | 0.000000e+00 |
| B | 2.323404e-15 | 1.000000e+00 | 1.000000e+00 | 3.010160e-01 | 2.384196e-39 |
| C | 1.486329e-79 | 1.000000e+00 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 7.380307e-177 | 3.010160e-01 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 0.000000e+00 | 2.384196e-39 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

분석 - Sement와 정상입금원금_B0M

earman 상관계수: 0.4073, p-value: 0.0000

skal-Wallis H-test

gment 등급 간에 당월 정상입금원금의 중앙값이나 분포 차이가 있는가?

skal-Wallis H-statistic: 399455.4783

alue: 0.0000

유의미한 차이가 있음. 귀무가설 기각됨

# 02 예측모델 구축

## 컬럼 선정

- 변수들의 비선형 관계를 고려해 **스피어만 상관계수가 0.4 이상인 변수들을 선별**하여 분석에 활용함.
- 선택된 변수들이 가설검정 결과와 논리적으로 일치하는지를 검토하여 **변수의 타당성을 확보**함.

## 학습모델 선정

- 대규모 정형 데이터를 기반으로 복잡한 비선형 관계를 학습하고 다중 클래스 분류 문제 해결을 위해 Logistic Regression(0.86), LightGBM(0.87), SGB(0.85), XGBoost(0.86) 모델을 적용함.
- Competition scoring 기준으로 성능을 비교한 결과, LightGBM이 가장 우수한 성능을 보여 최종 모델로 선정함.

**속도와 효율성**
- 대용량 데이터와 다수 변수에 대한 빠른 학습 및 예측 속도
- 멀티스레딩 지원 및 GPU 활용이 가능해 처리 시간 단축
- 리프 중심 트리 분할 방식으로 메모리 사용량 최적화

**결측치 및 범주형 변수 자동처리**
- 별도 전처리 없이 결측치 자동 처리 기능 제공
- 범주형 변수에 대한 최적의 분할 방식 자동 적용
- 데이터 전처리 과정 간소화로 모델 구축 속도 향상

**높은 예측 성능**
- Gradient Boosting 기반으로 복잡한 비선형 관계 포착
- 과적합 방지 메커니즘 내장으로 안정적 성능 유지

**특성 중요도 해석 가능**
- 각 변수의 중요도 시각화로 인사이트 도출 용이
- 세그먼트별 주요 영향 요인 식별 가능

# 02 예측모델 구축

## 모델 학습 방식

| 데이터셋 구성 | 변주형 변수 처리 |
| --- | --- |
| • 다수의 parquet 파일을 통합한 후, **ID 기준으로 병합**하여 단일 데이터셋 구성<br>• Segment 값을 타겟 변수로 설정하고, 관련 속성 변수들을 특징 변수로 분리 | • LightGBM의 **내장 범주형 변수 자동 인식 기능을 활용**하여 처리 효율성 향상<br>• 별도의 Label Encoding 없이도 범주형 변수 학습 가능 |
| **청크 단위 학습 적용** | **클래스 불균형 대응** |
| • 메모리 효율성 확보를 위해 데이터를 1,000개 단위로 **청크 분할**<br>• 순차적 누적 학습 방식을 적용하여 각 청크를 **순차적으로 모델에 학습**<br>• **청크 별 예측 결과의 정확도를 기준으로 평균화하여 전체 모델 성능 평가** | • 세그먼트 간 불균형 완화를 위해 모델 학습 시 **가중치 옵션 적용**<br>• ‘class_weight = balanced’를 설정하여 소수 Segment에 가중치 보정 수행<br>• 자동으로 클래스 빈도에 반비례하는 가중치를 부여하여 모델의 분류 편향 최소화 |

# 02 예측모델 구축

## 하이퍼파라미터 튜닝

**Optuna**를 활용한 최적화

- Optuna는 머신러닝 모델의 성능을 높이기 위한 **하이퍼파라미터 자동 탐색 프레임워크**
- 사람이 수동으로 조합을 실험하는 방식 대신 자동으로 최적에 가까운 조합을 탐색함

**Optuna** 선택 이유

- LightGBM의 다양한 하이퍼파라미터에 따른 **성능 변화 대응 필요**
- Grid Search나 Random Search는 탐색 공간이 너무 넓고 비효율적
- Optuna는 유망한 파라미터 조합 위주로 **지능적 탐색 수행하여 탐색 효율성과 성능 개선 효과를 동시에 달성**할 수 있어 적합함

## 모델 성능 결과

- Accuracy (예측된Segment와 실제 Segment간 일치 비율) : **0.84**
- 안정성 (모델 성능의 일관성 지표) : **±0.0128**
- Kaggle 점수 : **0.87567**

# 03 결론

## 과정 요약

| 1. 가설 기반 분석설계 | 2. 통계 및 머신러닝 분석 | 3.결과 |
|---|---|---|
| • **사전 가설 수립과 변수 검토**를 통해 데이터탐색의 신뢰성을 확보함.<br>• **고객 세그먼트별 특성**을 규명함.<br>• 세그먼트 구분에 중요한 역할을 하는 **핵심 차별화 변수를 도출**함 | • **Kruskal-Wallis H-test, 사후검정, 스피어만 상관분석**을 하여 LightGBM 모델을 통해 학습을 진행. | • LightGBM이 다른 모델 대비 최고 예측 정확도(0.87)를 기록함<br>• Optuna를 활용한 하이퍼파라미터 최적화로 **성능 안정성을 확보**함 |

# 03 결론

## 고객 페르소나

| Segment | 고객특성 | 전략 |
|---|---|---|
| A | • 거래규모가 크고, 장기 신뢰와 무결점 이력을 가진 최우수 고객 | • VIP 전용 프로그램<br>• 맞춤형 리워드<br>• 빠른 응대 채널 운영 등 |
| B | • 빈번한 거래와 높은 상환 의지<br>• 금융상품에 적극적 | • 추천 리워드<br>• 장기 고객 감사<br>• 신상품 정보 우선 제공 등 |
| C | • 평균적 사용과 안정적 상환<br>• 전형적인 메인스트림 고객 | • 대중 마케팅<br>• 신규 서비스 체험 기회 제공 등 |
| D | • 사용빈도 및 금액 모두 낮고 반응성 낮음 | • 관심 기반 콘텐츠 마케팅<br>• 실속형 이벤트 메시지 등 |
| E | • 거래 중단 또는 미미하고 연체 이력은 크지 않으나 관계 단절 위험이 높음 | • 재방문 쿠폰<br>• 감성 메시지<br>• 설문 기반 리마케팅 등 |

# 03 결론

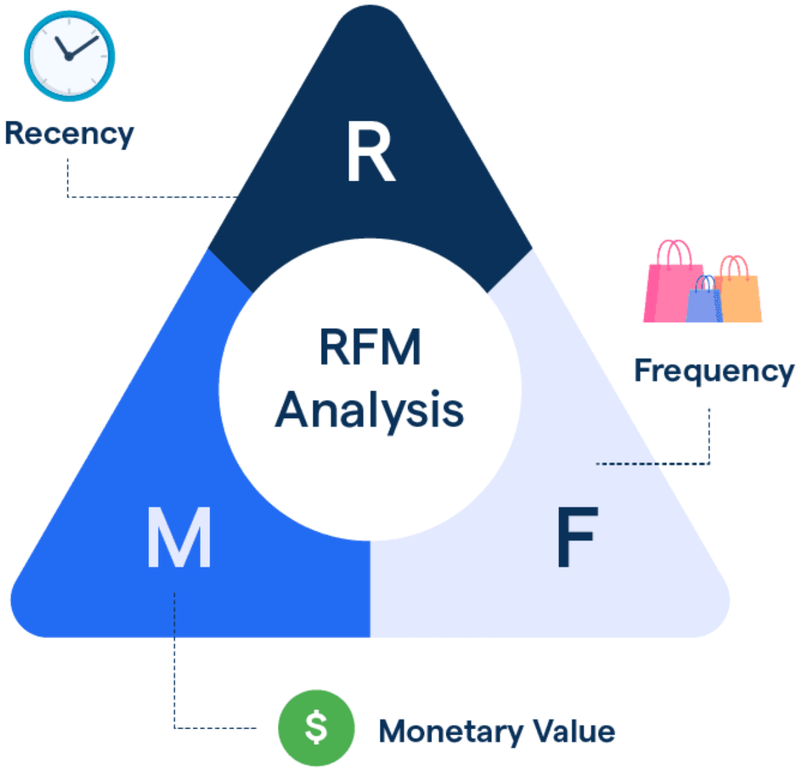
## 추가분석 : RFM

**분석 목적**

- 기존 Segment 분류를 보완하기 위해 고객의 **거래 행동 데이터를 기반**으로 RFM 분석을 실시함.
- 이를 통해 거래 시점(Recency), 거래 빈도(Frequency), 거래 금액(Monetary) 지표를 통합적으로 고려한 고객 세분화를 시도함.
- 실제 행동에 기반한 세분화는 마케팅, 리텐션 전략 등에서 중요한 접근이라고 판단함.

**RFM 등급 분류 기준**

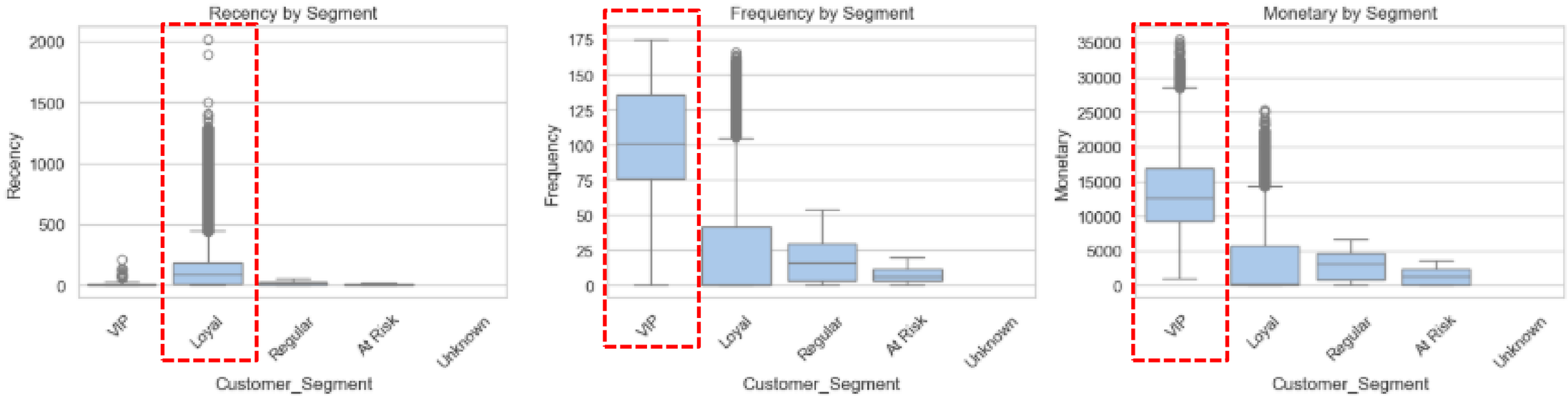
- **VIP**: 최근 거래 + 고빈도 + 고금액 → 핵심 우수고객
- **Loyal**: 거래 빈도는 높지만 금액은 중간 수준인 고객
  → 장기 유지 고객 & 충성도가 높은 고객
- **At Risk**: 최근 거래량이 낮은 이탈 위험군
  → 리텐션 캠페인, 재활성화 마케팅의 주요 타겟이 될 수 있음
- **Regular**: 평균 수준의 일반 고객
  → 잠재 성장 가능성은 낮으나 안정성을 담당



# 03 결론

## 추가분석 : RFM

분석결과(1)

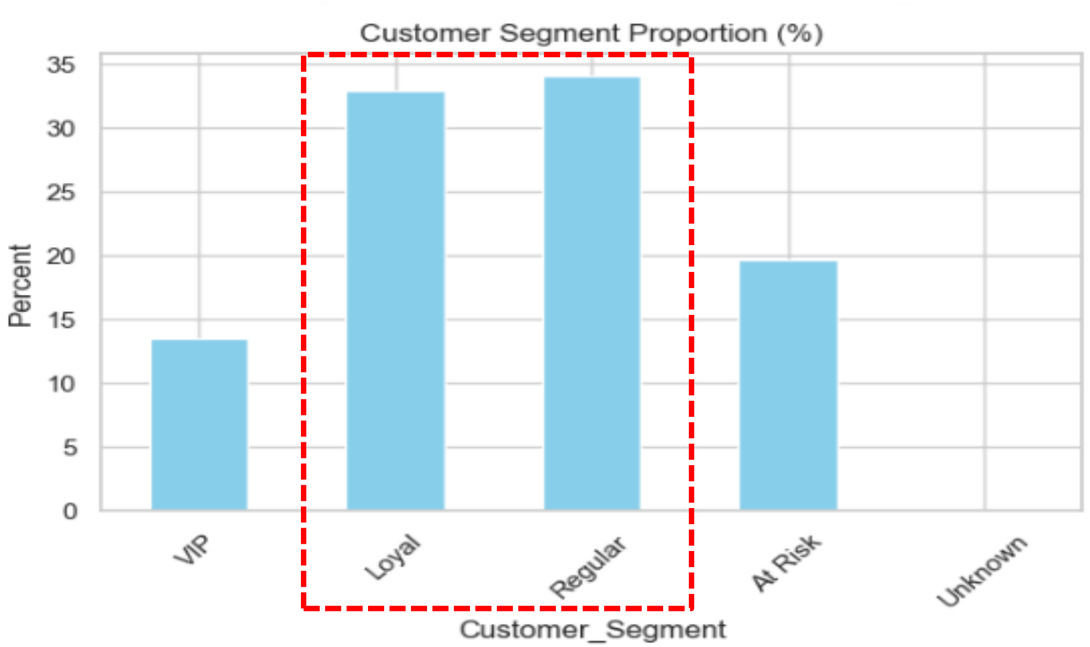



- RFM 분석 결과, VIP 고객은 최근성(Recency)이 낮고 구매 빈도(Frequency)와 금액(Monetary)이 모두 높은 수준으로 나타나, 기업에 가장 높은 가치를 제공하는 고객군으로 식별됨.
- 반면 Loyal 및 Regular 고객은 상대적으로 구매 활동이 점진적으로 감소하는 양상을 보임.
- At Risk 고객은 구매 이력이 미미하거나 장기간 활동이 없어 이탈 위험이 높고, 이에 따른 재활성화 및 관리 전략의 필요성이 제기됨.

# 03 결론

## 추가분석 : RFM

분석결과(2)




| Customer_Segment | |
|---|---|
| Regular | 131086 |
| Loyal | 126477 |
| At Risk | 75552 |
| VIP | 51615 |

- 전체 고객의 약 65%는 Regular (34%) 및 Loyal (32%) 고객으로 안정적인 매출 기반을 형성함.
- VIP 고객 (14%)은 규모는 작지만 매출 기여도가 가장 높은 핵심 고객군으로 고도화된 유지 전략이 필요함.
- At Risk 고객(20%)은 이탈 가능성이 높아 재활성화를 위한 타겟 마케팅이 시급함.
- RFM 지표를 모델링에 통합했다면 고객의 가치 기반 세분화가 가능해져 예측 성능 향상에 기여했을 수 있음.

# 결론

01 가설 기반 설계와 통계검정을 통해 변수의 논리성과 신뢰도를 확보함

02 머신러닝 이전 단계에서 변수 선택의 완성도를 높임

03 EDA 결과와 모델 입력 변수 간 연계, 피처 중요도 기반 전략 제안 등의 보완이 필요함

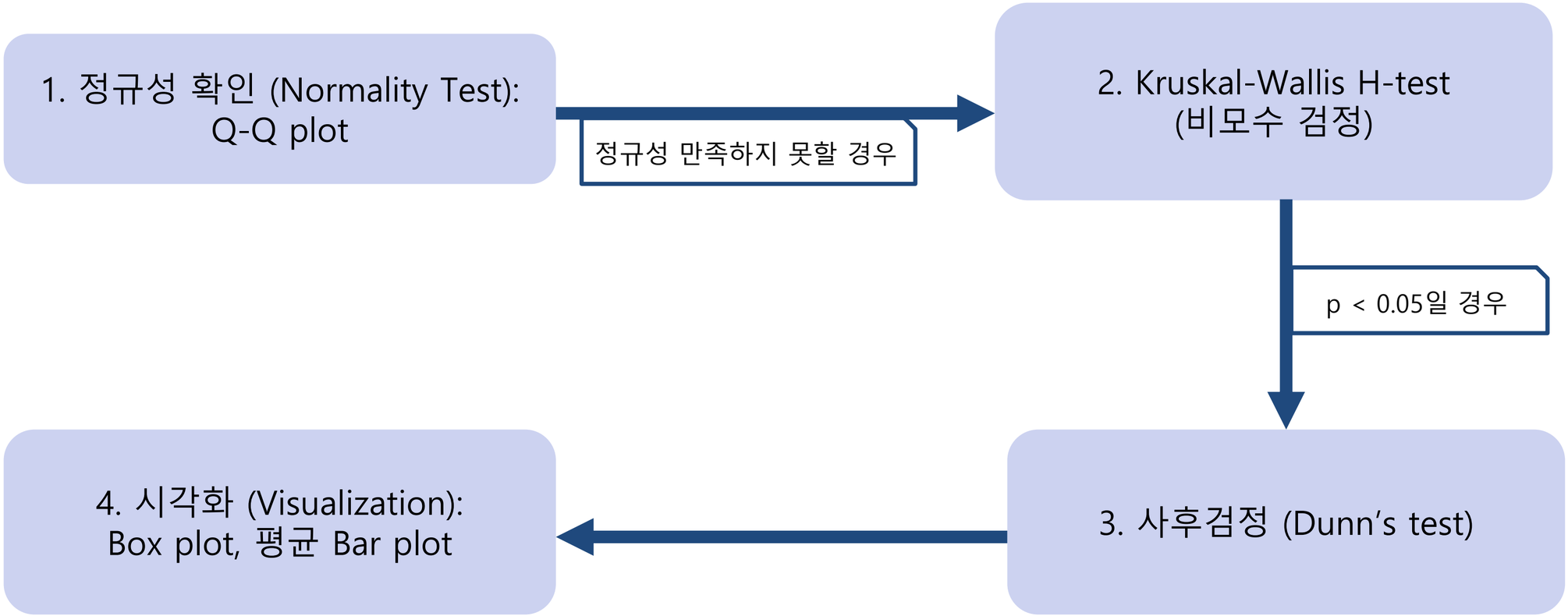
04 이를 통해 다음 분석에서는 모델 해석력과 전략 활용 가능성을 더욱 높일 수 있을 것으로 기대함

# 감사합니다

# 04 부록1: 가설 검정과정

## 가설검정 과정

- 가설검정 과정은 다음의 4단계를 거침

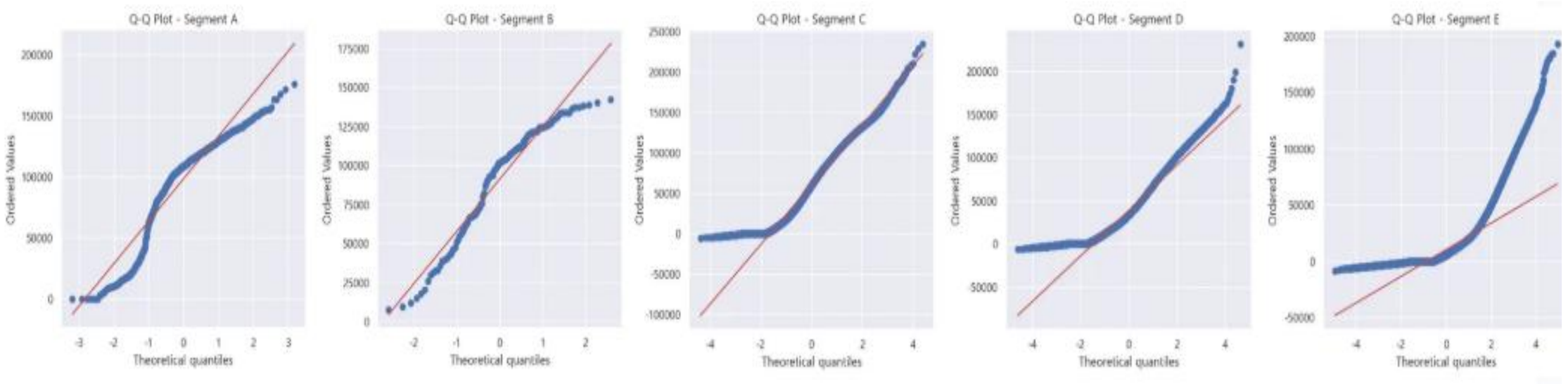


# 04 부록1: 가설 검정과정

**가설 1.**

최근 3개월간 신용/체크카드 이용금액이 높을수록 고객 가치가 높을 것이다

1. 정규성 검정




- 이론선(발간색 선)은 정규분포를 따를 경우 데이터가 위치해야 할 기준 직선임.
- Q-Q Plot 관찰 결과 모든 그룹에서 S자 형태로 만곡되어 있고, 양 끝 부분에서 실제 관측값이 이론선으로부터 현저히 이탈함.
- 해당 변수는 전 그룹에서 정규성 가정을 충족하지 않으며, 비모수 검정 등의 대안적 접근이 요구됨.

# 04 부록1: 가설 검정과정

**가설 1.**

최근 3개월간 신용/체크카드 이용금액이 높을수록 고객 가치가 높을 것이다

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 56799.8862
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.000000e+00 | 2.470422e-27 | 1.898502e-82 | 0.000000e+00 |
| B | 1.000000e+00 | 1.000000e+00 | 1.042800e-04 | 2.883591e-13 | 1.823854e-104 |
| C | 2.470422e-27 | 1.042800e-04 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 1.898502e-82 | 2.883591e-13 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 0.000000e+00 | 1.823854e-104 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

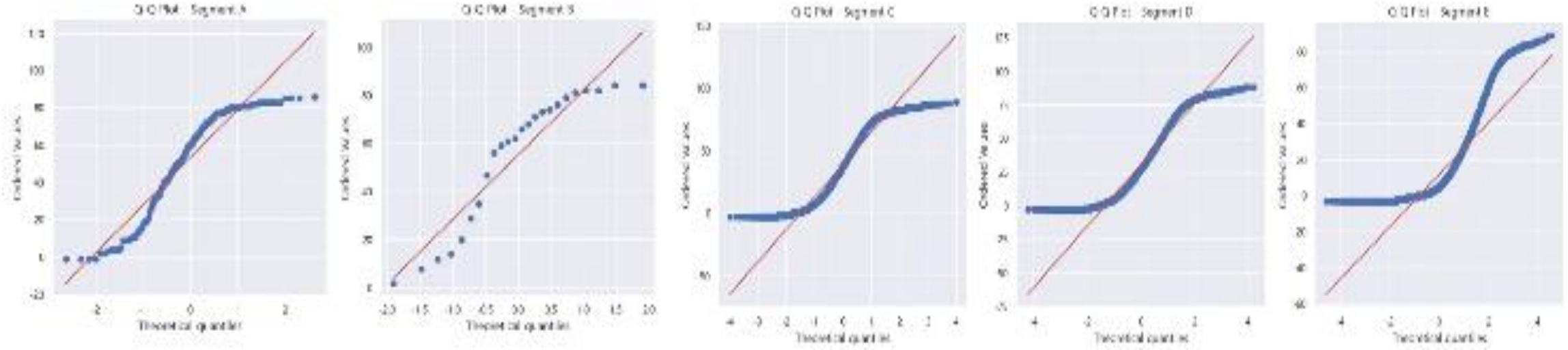
- Segment A와 B 간 이용금액_R3M_신용체크(최근 3개월 동안의 신용/체크카드 이용금액)에 대한 차이가 유의하지 않음(p=1)
- 나머지 그룹들 간 이용가능카드수_신용에 대한 차이는 유의함(p<0.05)

# 04 부록1: 가설 검정과정

**가설 2.**

카드 사용 빈도가 증가할수록 고객 가치는 높아질 것이다.

1. 정규성 검정



- 종속변수: 총 거래 빈도 (= 이용건수_일시불_B0M + 이용건수_할부_B0M)
- 총 거래 빈도는 '이산형 수치형'이나 값의 범위가 충분히 커서 해당 컬럼에 대한 정규성을 확인함.
- 확인 결과 가설 1과 동일하게 모든 그룹에서 '총거래빈도'는 정규성을 만족하지 못함.

# 04 부록1: 가설 검정과정

**가설 2.**

카드 사용 빈도가 증가할수록 고객 가치는 높아질 것이다.

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 611194.6708
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.000000e+00 | 1.630311e-04 | 6.949824e-11 | 2.743323e-72 |
| B | 1.000000e+00 | 1.000000e+00 | 5.923598e-01 | 4.168155e-02 | 6.314206e-12 |
| C | 1.630311e-04 | 5.923598e-01 | 1.000000e+00 | 4.497291e-136 | 0.000000e+00 |
| D | 6.949824e-11 | 4.168155e-02 | 4.497291e-136 | 1.000000e+00 | 0.000000e+00 |
| E | 2.743323e-72 | 6.314206e-12 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

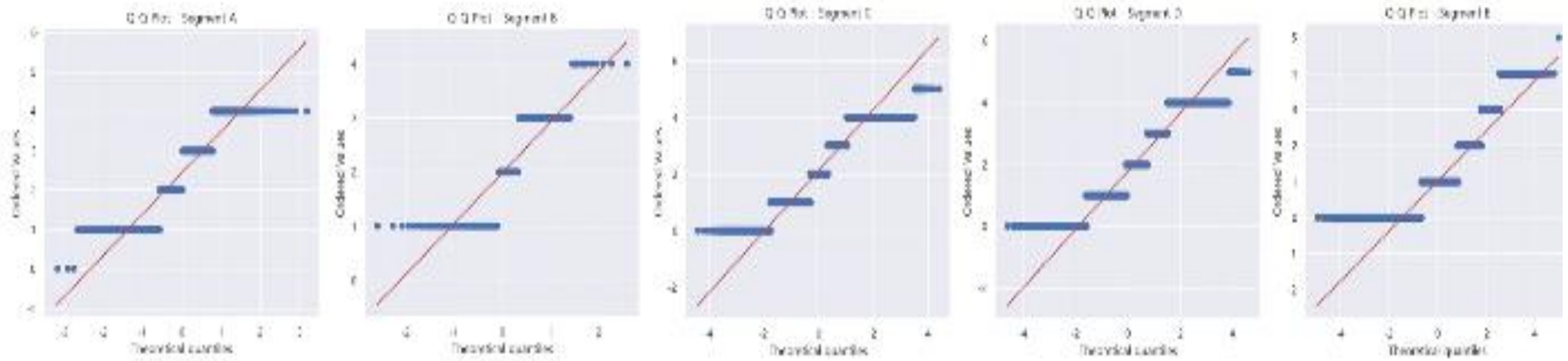
- Segment A와 B, B와 C간 카드 사용 빈도에 대한 차이가 유의하지 않음(p>0.05)
- 나머지 그룹 간 카드 사용 빈도에 대한 차이는 유의함(p<0.05)

# 04 부록1: 가설 검정과정

**가설 3.**

실제 사용 중인 신용카드 수가 많을 수록 고객 가치가 높을 것이다

1. 정규성 검정



- Q-Q Plot 상 데이터들이 불연속적인 수평선 형태로 나타남.
- 해당 형태는 종속변수인 이용카드수_신용 변수가 가지는 이산형 특성에 기인한 것으로 보임.
- 모든 그룹에서 종속변수는 정규성을 충족하지 않음.

# 04 부록1: 가설 검정과정

**가설 3.**

실제 사용 중인 신용카드 수가 많을 수록 고객 가치가 높을 것이다

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 281950.7895
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 3.213828e-03 | 1.768081e-14 | 5.511538e-48 | 0.000000e+00 |
| B | 3.213828e-03 | 1.000000e+00 | 1.000000e+00 | 6.953753e-01 | 9.792257e-29 |
| C | 1.768081e-14 | 1.000000e+00 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 5.511538e-48 | 6.953753e-01 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 0.000000e+00 | 9.792257e-29 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

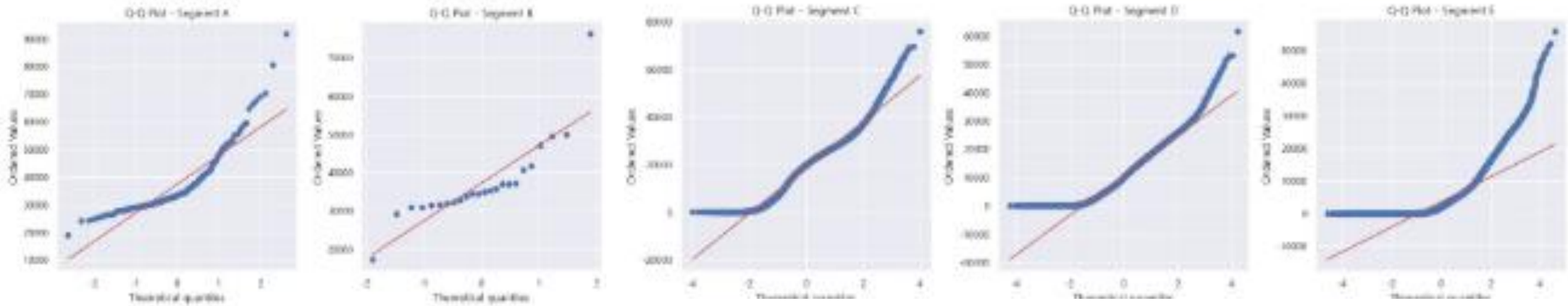
- Segment B와 C, B와 D간 '이용카드수_신용'에 대한 차이가 통계적으로 유의하지 않음 (p>0.05)
- 나머지 그룹들 간 '이용카드수_신용'에 대한 차이는 유의함 (p<0.05)

# 04 부록1: 가설 검정과정

**가설 4.**

월 청구금액이 높을수록 Segment등급도 높을 것이다

1. 정규성 검정



- 앞선 가설들과 동일한 패턴이 관찰됨.
- 모든 그룹에서 종속변수는 정규성을 충족하지 않음.

# 04 부록1: 가설 검정과정

**가설 4.**

월 청구금액이 높을수록 Segment등급도 높을 것이다

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 93342.0572
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.000000e+00 | 1.215525e-07 | 2.639790e-28 | 1.059118e-135 |
| B | 1.000000e+00 | 1.000000e+00 | 2.895297e-01 | 1.597142e-04 | 1.125649e-20 |
| C | 1.215525e-07 | 2.895297e-01 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 2.639790e-28 | 1.597142e-04 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 1.059118e-135 | 1.125649e-20 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

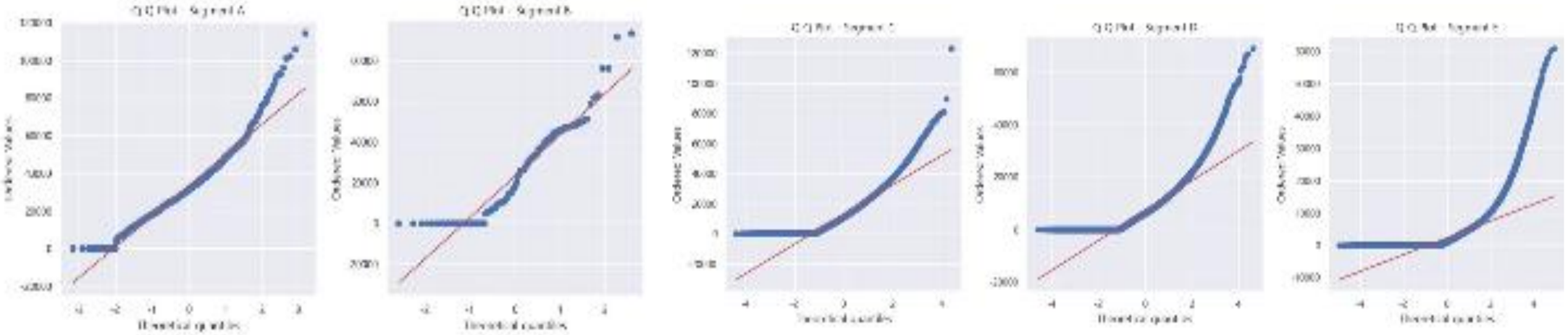
- Segment A와 B, B와 C간 '정상청구원금_B0M(당월 정상청구원금)'에 대한 차이가 통계적으로 유의하지 않음(p>0.05)
- 나머지 그룹간 '정상청구원금_B0M'에 대한 차이는 유의함(p>0.05)

# 04 부록1: 가설 검정과정

**가설 5.**

정상입금원금이 높을수록 고객이 상위 등급(Segment A 또는 B)에 속할 가능성이 높을 것이다

1. 정규성 검정



- 앞선 가설들과 동일한 패턴이 관찰됨.
- 모든 그룹에 대하여 종속변수인 '정상입금원금_B0M(당월 정상입금원금)'은 정규성을 만족하지 못함.

# 04 부록1: 가설 검정과정

**가설 5.**

정상입금원금이 높을수록 고객이 상위 등급(Segment A 또는 B)에 속할 가능성이 높을 것이다

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 399455.4783
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 2.323404e-15 | 1.486329e-79 | 7.380307e-177 | 0.000000e+00 |
| B | 2.323404e-15 | 1.000000e+00 | 1.000000e+00 | 3.010160e-01 | 2.384196e-39 |
| C | 1.486329e-79 | 1.000000e+00 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 7.380307e-177 | 3.010160e-01 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 0.000000e+00 | 2.384196e-39 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

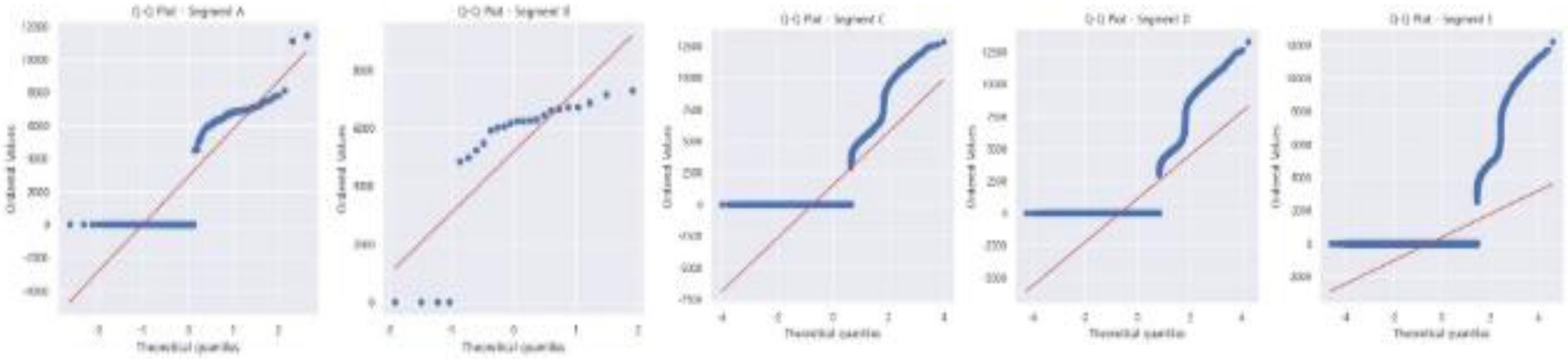
- Segment B와 C, B와 D간 '정상입금원금_B0M(당월 정상입금원금)'에 대한 차이가 통계적으로 유의하지 않음(p>0.05)
- 나머지 그룹간 '정상입금원금_B0M'에 대한 차이는 유의함(p>0.05)

# 04 부록1: 가설 검정과정

**가설 6.**

카드사로부터 대출 금액이 클수록 고객의 중요도가 높을 것이다.

1. 정규성 검정



- 해당 가설의 종속변수는 정상청구원금_B0M(당월정상청구원금)과 할부이용금액이나 정상청구원금에 대해서는 가설 4에서 이미 검정을 완료했으므로, 가설 6에서는 할부이용금액에 대한 검정을 진행함.
- Q-Q plot 모두 정규분포 직선에서 명확히 벗어나는 형태를 보여 모든 그룹에 대하여 '할부이용금액'은 정규성을 만족하지 못함.

# 04 부록1: 가설 검정과정

**가설 6.**

카드사로부터 대출 금액이 클수록 고객의 중요도가 높을 것이다.

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 16920.4486
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 3.083985e-09 | 1.864258e-17 | 4.085953e-29 | 2.836542e-64 |
| B | 3.083985e-09 | 1.000000e+00 | 4.257771e-23 | 8.277081e-28 | 1.943415e-39 |
| C | 1.864258e-17 | 4.257771e-23 | 1.000000e+00 | 1.811649e-144 | 0.000000e+00 |
| D | 4.085953e-29 | 8.277081e-28 | 1.811649e-144 | 1.000000e+00 | 0.000000e+00 |
| E | 2.836542e-64 | 1.943415e-39 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

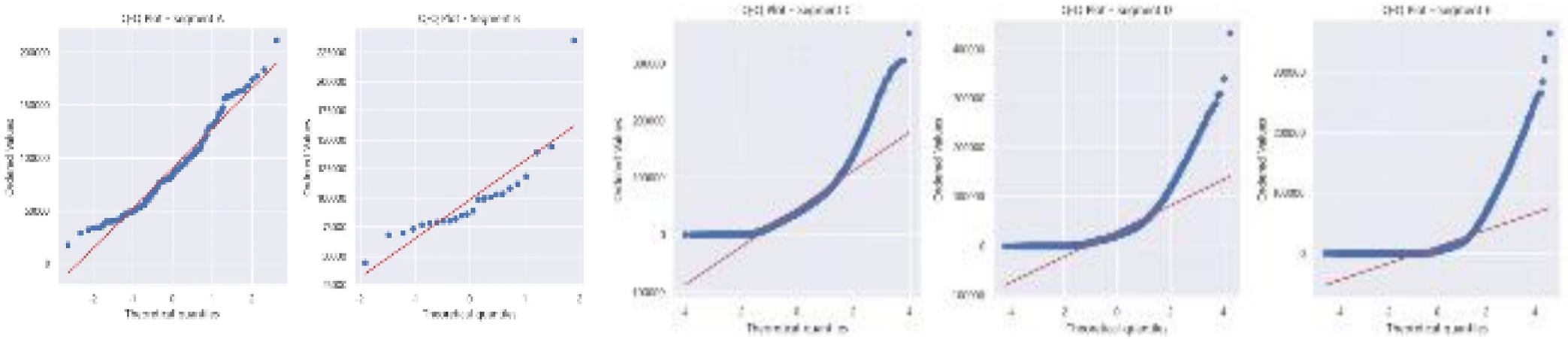
- 모든 그룹들 간 '할부이용금액'에 대한 차이는 통계적으로 유의함(p<0.05)

# 04 부록1: 가설 검정과정

**가설 7.**

연체 잔액이 높은 고객은 주로 A, B등급의 핵심 고객군에 집중될 것이다 (총잔액)

1. 정규성 검정



- Segment A와 B은 정규분포 직선에서 약하게 벗어난 모습을 보인 반면, Segment C, D, E에서는 양쪽에서 크게 벗어나는 형태를 보임.
- 즉, 모든 그룹에 대하여 '총잔액'은 정규성을 만족하지 못함.

# 04 부록1: 가설 검정과정

**가설 7.**

연체 잔액이 높은 고객은 주로 A, B등급의 핵심 고객군에 집중될 것이다 (총잔액)

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 76048.1284
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.000000e+00 | 5.222162e-10 | 2.984821e-25 | 1.326296e-118 |
| B | 1.000000e+00 | 1.000000e+00 | 5.600054e-02 | 1.558290e-04 | 4.016131e-19 |
| C | 5.222162e-10 | 5.600054e-02 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 2.984821e-25 | 1.558290e-04 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 1.326296e-118 | 4.016131e-19 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

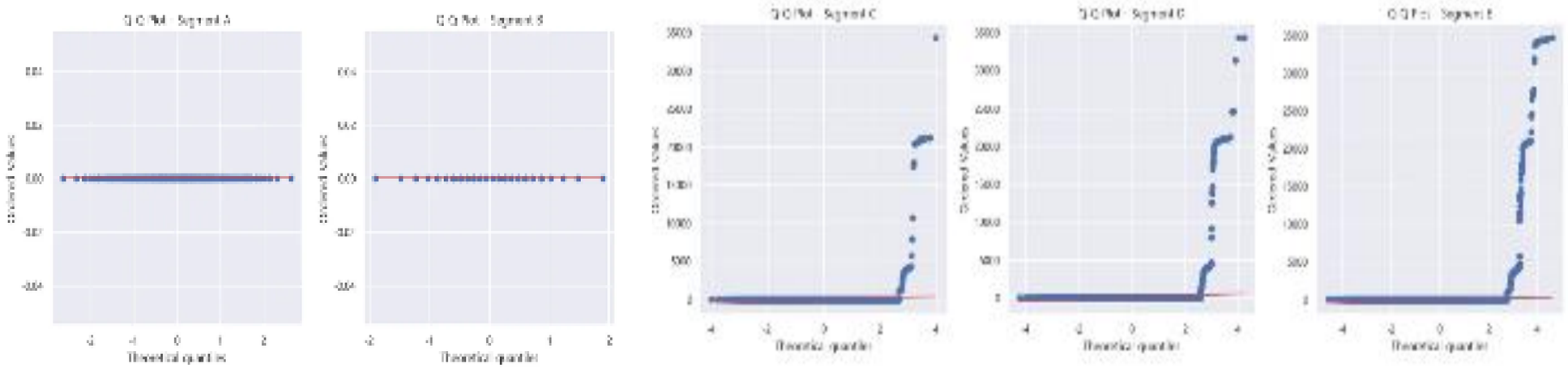
- Segment A와 B간 총잔액에 대한 차이가 통계적으로 유의하지 않음(p>0.05)
- 나머지 그룹 간 총잔액에 대한 차이는 통계적으로 유의함(p<0.05)

# 04 부록1: 가설 검정과정

**가설 7.**

연체 잔액이 높은 고객은 주로 A, B등급의 핵심 고객군에 집중될 것이다 (연체잔액)

1. 정규성 검정



- Segment A와 B의 경우 거의 모든 점이 0에 집중되어 있고 수평선 위에 존재함. 즉 정규성 판단 자체가 무의미함.
- Segment C, D, E의 경우 Q-Q plot 내 직선에서 크게 벗어나는 것을 확인할 수 있음. 즉 모든 그룹에서 정규성을 위배함.

# 04 부록1: 가설 검정과정

**가설 7.**

연체 잔액이 높은 고객은 주로 A, B등급의 핵심 고객군에 집중될 것이다 (연체잔액)

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 57.4109
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.0 | 1.0 | 1.00000 | 1.000000e+00 | 1.000000e+00 |
| B | 1.0 | 1.0 | 1.00000 | 1.000000e+00 | 1.000000e+00 |
| C | 1.0 | 1.0 | 1.00000 | 3.472041e-02 | 1.000000e+00 |
| D | 1.0 | 1.0 | 0.03472 | 1.000000e+00 | 6.362873e-13 |
| E | 1.0 | 1.0 | 1.00000 | 6.362873e-13 | 1.000000e+00 |

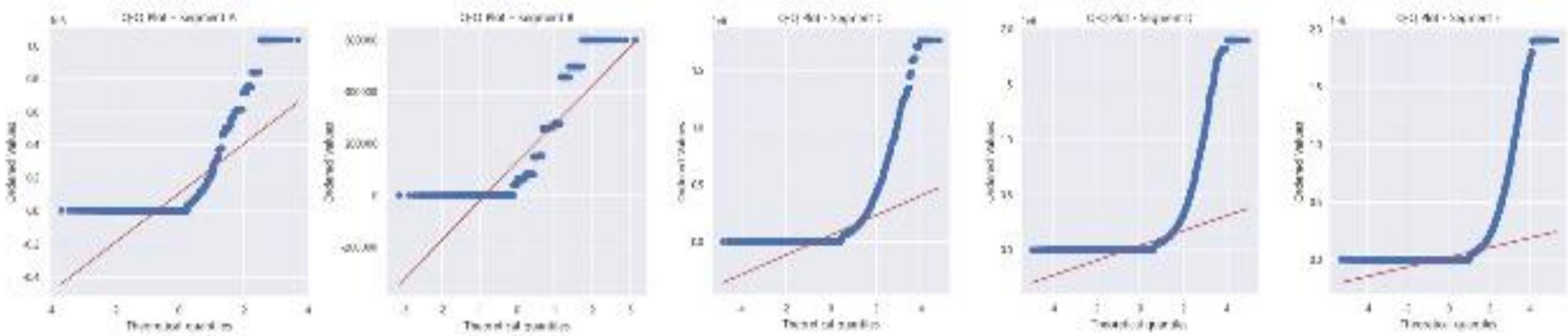
- Segment C와 D, D와 E를 제외한 나머지 그룹 간 연체잔액 차이는 통계적으로 유의하지 않음.
- Segment C와 D, D와 E간 연체잔액 차이만 통계적으로 유의함.

# 04 부록1: 가설 검정과정

**가설 8.**

세그먼트 E는 재무/소비 관련 지표에서 평균적인 특성을 보이는 세그먼트일 것이다
(카드론 이용금액 누적)

1. 정규성 검정



- 플롯 모두 오른쪽으로 갈수록 정규분포 직선(red line)에서 위로 벗어난 형태를 보이고 있음.
- 즉 모든 그룹에 대하여 '카드론이용금액_누적'은 정규성을 만족하지 못함.

# 04 부록1: 가설 검정과정

**가설 8.**

세그먼트 E는 재무/소비 관련 지표에서 평균적인 특성을 보이는 세그먼트일 것이다
(카드론 이용금액 누적)

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 277622.6499
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.129372e-10 | 1.836057e-31 | 8.006862e-100 | 1.861399e-283 |
| B | 1.129372e-10 | 1.000000e+00 | 2.117444e-31 | 4.221611e-53 | 2.779288e-98 |
| C | 1.836057e-31 | 2.117444e-31 | 1.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| D | 8.006862e-100 | 4.221611e-53 | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 |
| E | 1.861399e-283 | 2.779288e-98 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

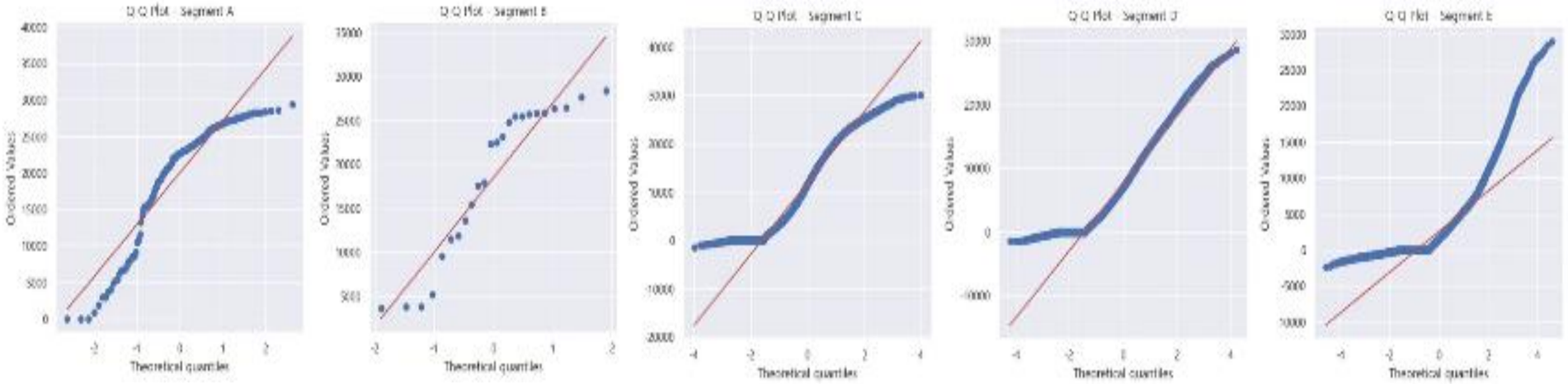
- 모든 그룹들 간 '카드론이용금액_누적'에 대한 차이는 통계적으로 유의함(p<0.05)

# 04 부록1: 가설 검정과정

**가설 8.**

세그먼트 E는 재무/소비 관련 지표에서 평균적인 특성을 보이는 세그먼트일 것이다
(이용금액_일시불)

1. 정규성 검정




- 플롯 모두 정규분포 직선인 redline에서 벗어난 형태를 보이고 있음.
- 즉 모든 그룹에 대하여 '이용금액_일시불'은 정규성을 만족하지 못한다고 해석할 수 있음.

# 04 부록1: 가설 검정과정

**가설 8.**

세그먼트 E는 재무/소비 관련 지표에서 평균적인 특성을 보이는 세그먼트일 것이다 (이용금액_일시불)

2. Kruskal-Wallis H-test (비모수 검정)

```
Kruskal-Wallis H-statistic: 78628.2552
p-value: 0.0000
→ 유의미한 차이가 있음. 귀무가설 기각됨
```

- H-test 결과 p<0.05로 귀무가설이 기각됨.

3. 사후검정 (Dunn's test)

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 1.000000e+00 | 1.000000e+00 | 3.576230e-07 | 8.063167e-20 | 2.660433e-109 |
| B | 1.000000e+00 | 1.000000e+00 | 6.300725e-01 | 8.449638e-03 | 8.657404e-16 |
| C | 3.576230e-07 | 6.300725e-01 | 1.000000e+00 | 2.735913e-309 | 0.000000e+00 |
| D | 8.063167e-20 | 8.449638e-03 | 2.735913e-309 | 1.000000e+00 | 0.000000e+00 |
| E | 2.660433e-109 | 8.657404e-16 | 0.000000e+00 | 0.000000e+00 | 1.000000e+00 |

- Segment A와 B, B와 C를 제외한 나머지 그룹들 간 이용금액_일시불 차이는 통계적으로 유의하지 않음(p>0.05)
- 나머지 그룹들 간 이용금액_일시불에 대한 차이는 통계적으로 유의함(p<0.05)