데이터 전처리 결과서

I. 서론

1. 배경 및 목적

- 데이터 분석의 정확성과 효율성을 높여 더 신뢰할 수 있는 결과를 얻기 위해 수행
- 결측치 및 이상치 처리, 데이터 정규화, 인코딩 등을 통해 데이터의 품질을 개선

2. 프로젝트 개요

- 프로젝트 주제: 헬스장 이탈률 분석 및 예측 모델 개발

3. 데이터 개요

- 출처: https://www.kaggle.com/code/ellanihill/customer-churn-analysis/notebook
- 총 샘플 수: 4002개, 컬럼 수: 14개, 결측치: 0개
- 주요 변수(컬럼) 목록 및 특성

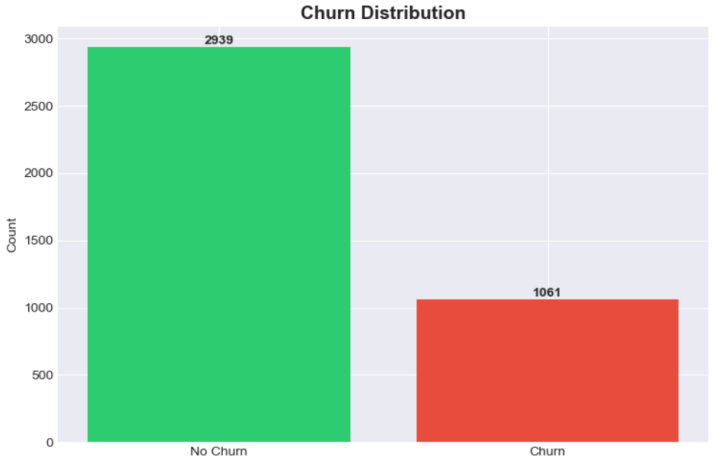
번호	컬럼명	설명	데이터 타입	데이터 유형
1	gender	성별	int	범주형
2	Near_Location	거주지 인근 여부	int	범주형
3	Partner	파트너 회원 여부	int	범주형
4	Promo_friends	친구 추천 프로모션	int	범주형
5	Phone	연락처 등록 여부	int	범주형
6	Contract_period	계약 기간(1/6/12개월)	int	범주형
7	Group_visits	그룹 수업 참여	int	범주형
8	Age	나이	int	수치형
9	Avg_additional_charges_total	평균 추가 요금	float	수치형
10	Month_to_end_contract	계약 만료 잔여 개월	float	수치형
11	Lifetime	회원 활동 기간(개월)	int	수치형
12	Avg_class_frequency_total	전체 수업 빈도	float	수치형

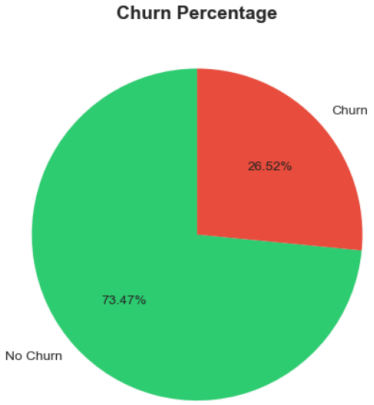
13	Avg_class_frequency_current_month	최근 월 수업 빈도	float	수치형
14	Churn	이탈 여부	int	타겟(범주형)

II. 전처리 수행 과정

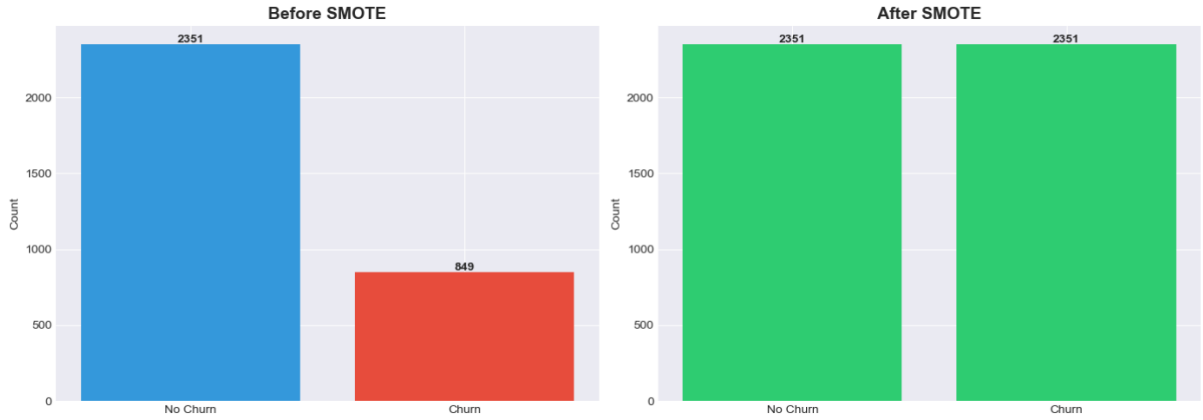
3. 데이터 이해 및 초기 분석 (EDA 결과 요약)

- 타겟 변수(Churn) 분석 (클래스 분포)

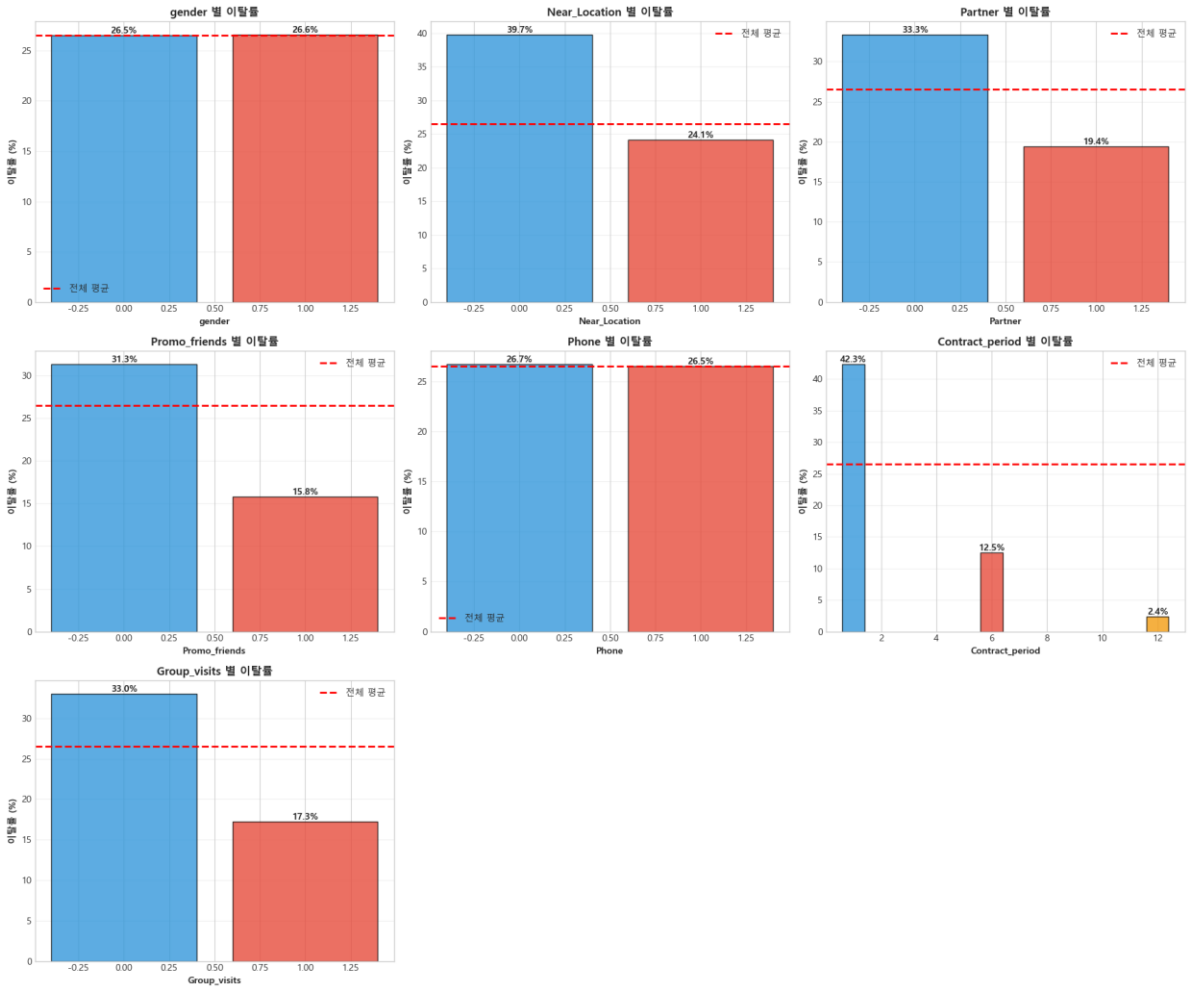




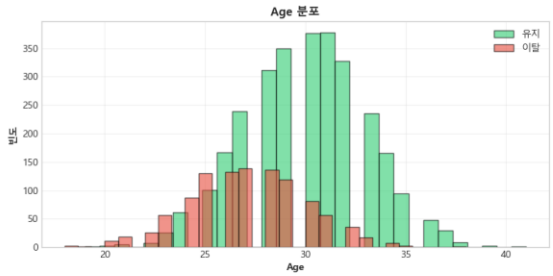
- 타겟 클래스 불균형 해결

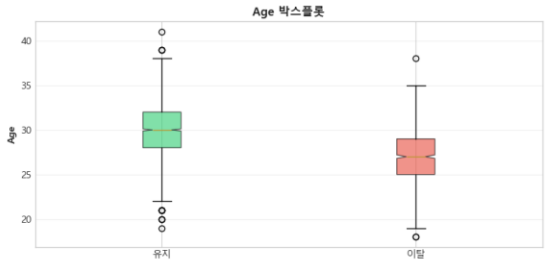


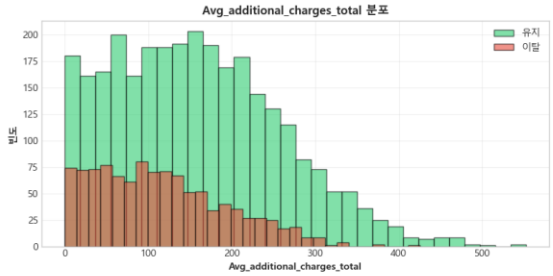
- 범주형 변수 분석

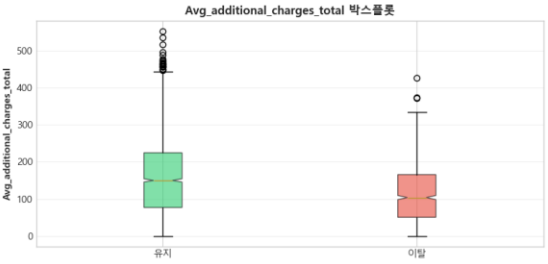

gender 별 이탈률
26.5%
26.6%
전체 평균
이탈률 (%)
gender
Near_Location 별 이탈률
39.7%
24.1%
전체 평균
이탈률 (%)
Near_Location
Partner 별 이탈률
33.3%
19.4%
전체 평균
이탈률 (%)
Partner
Promo_friends 별 이탈률
31.3%
15.8%
전체 평균
이탈률 (%)
Promo_friends
Phone 별 이탈률
26.7%
26.5%
전체 평균
이탈률 (%)
Phone
Contract_period 별 이탈률
42.3%
12.5%
2.4%
전체 평균
이탈률 (%)
Contract_period
Group_visits 별 이탈률
33.0%
17.3%
전체 평균
이탈률 (%)
Group_visits

- 수치형 변수 분석

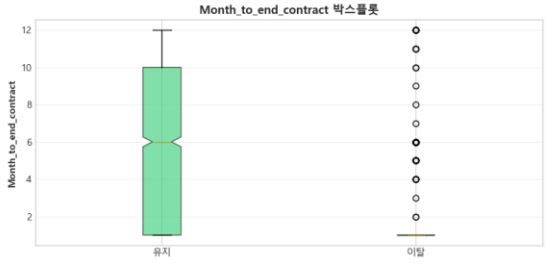

Age 분포
유지
이탈
빈도
Age

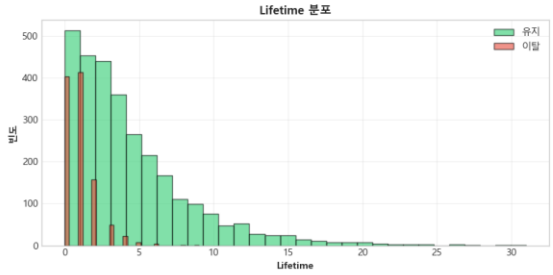

Age 박스플롯
Age
유지
이탈

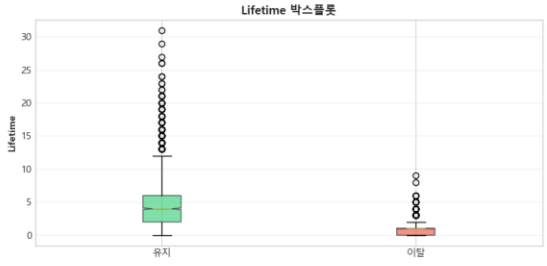

Avg_additional_charges_total 분포
유지
이탈
빈도
Avg_additional_charges_total

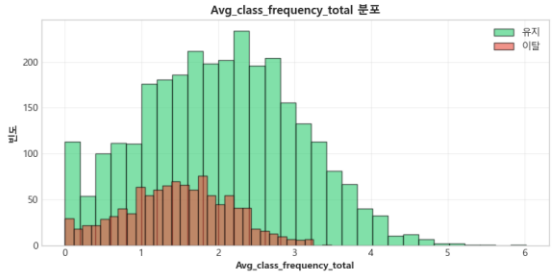

Avg_additional_charges_total 박스플롯
Avg_additional_charges_total
유지
이탈

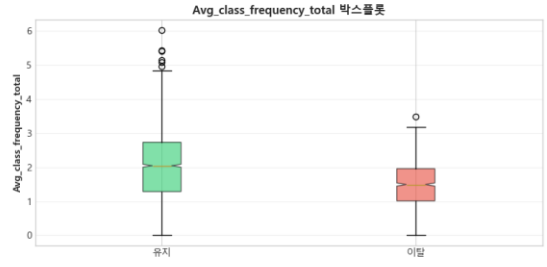

Month_to_end_contract 분포
유지
이탈
빈도
Month_to_end_contract

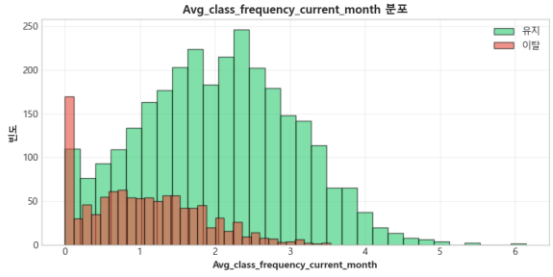

Month_to_end_contract 박스플롯
Month_to_end_contract
유지
이탈

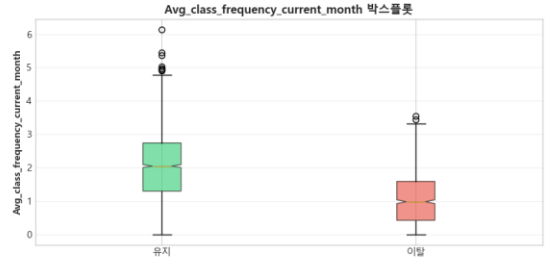

Lifetime 분포
유지
이탈
빈도
Lifetime


Lifetime 박스플롯
Lifetime
유지
이탈

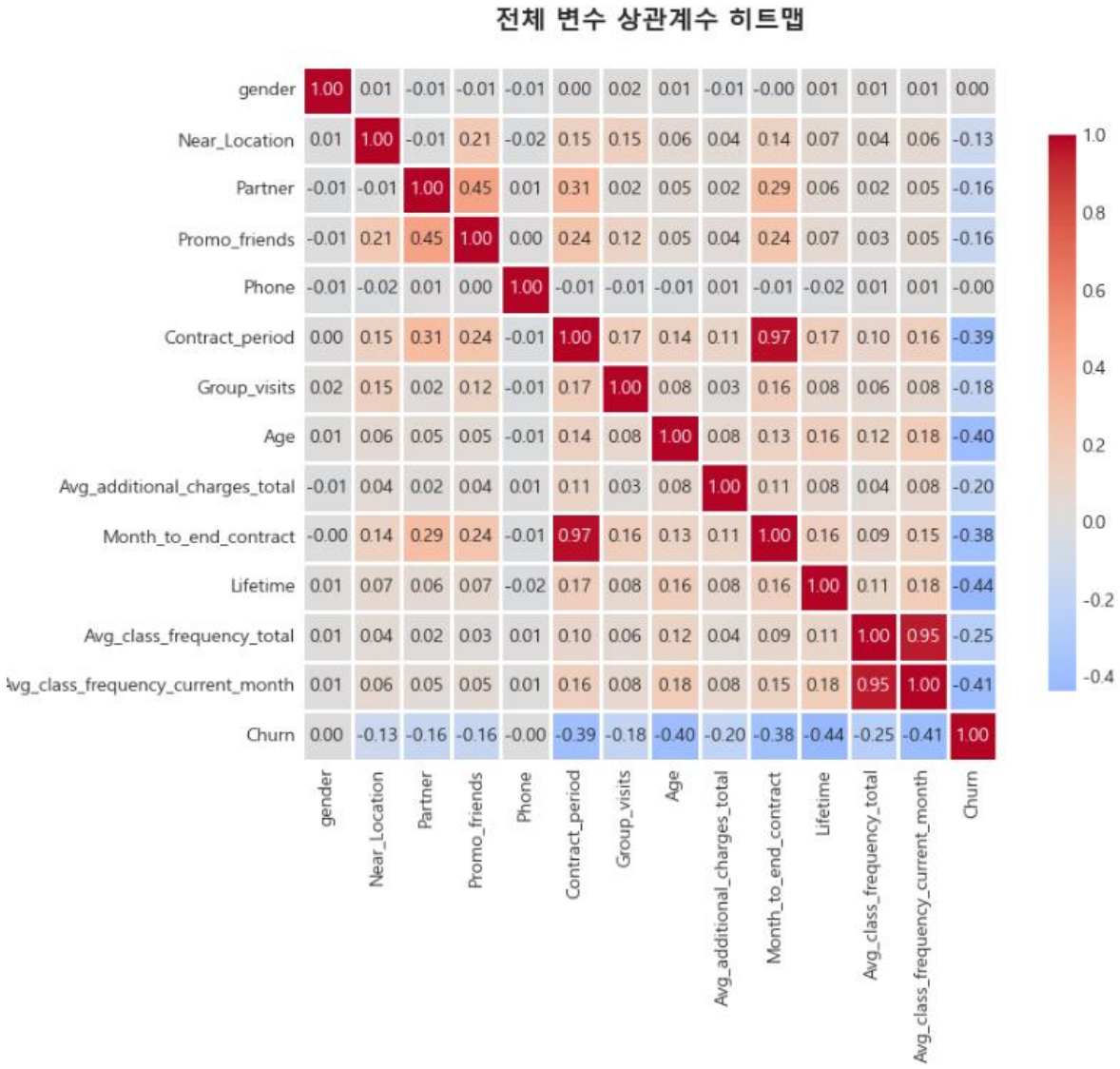

Avg_class_frequency_total 분포
유지
이탈
빈도
Avg_class_frequency_total


Avg_class_frequency_total 박스플롯
Avg_class_frequency_total
유지
이탈


Avg_class_frequency_current_month 분포
유지
이탈
빈도
Avg_class_frequency_current_month


Avg_class_frequency_current_month 박스플롯
Avg_class_frequency_current_month
유지
이탈

- 전체 변수 상관계수 히트맵



- 데이터 품질: 결측치 0개 => 결측치 처리 불필요

4. 데이터 정제 및 변환

- 데이터 스케일링: StandardScaler
- 데이터 샘플링: SMOTE
- 파생변수 생성

번호	컬럼명	설명	데이터 타입	데이터 유형
1	LIfetime_per_month	계약 대비 실제 이용 효율	float	수치형
2	Is_New_Member	신규 회원 플래그	int	범주형
3	Is_Long_Member	장기 회원 플래그	int	범주형
4	Class_Engagement	누적 수업 참여도	float	수치형
5	Recent_Activity	최근 활동 변화율	float	수치형
6	Contract_Completioin	계약 진행도	float	범주형
7	Long_Contract	장기 계약 플래그	int	범주형
8	Cost_per_Visit	방문당 평균 지출	float	수치형
9	High_Spender	고지출 회원 플래그	int	범주형
10	Engagement_Score	종합 참여도 점수	int	수치형
11	Churn_Risk	복합 리스크 점수	int	수치형

III. 전처리 결과 및 분석

5. 전처리 전/후 비교

- 타겟 클래스 불균형 해결

SMOTE 적용 전

클래스 분포: 유지 2,255명, 이탈 946명

불균형 비율: 2.33:1

문제점: 소수 클래스(이탈) 패턴을 제대로 학습하지 못함

SMOTE 적용 후

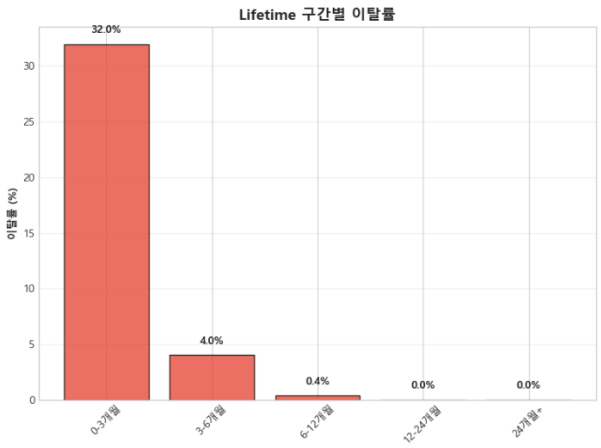
클래스 분포: 유지 2,255명, 이탈 2,255명

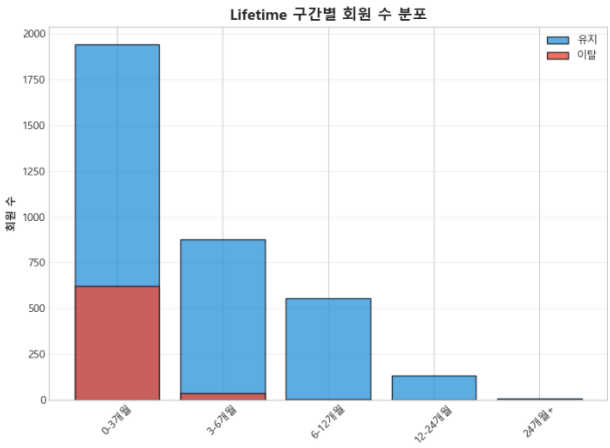
균형 비율: 1:1

효과: 모델이 이탈 고객 패턴을 더욱 정확하게 학습함

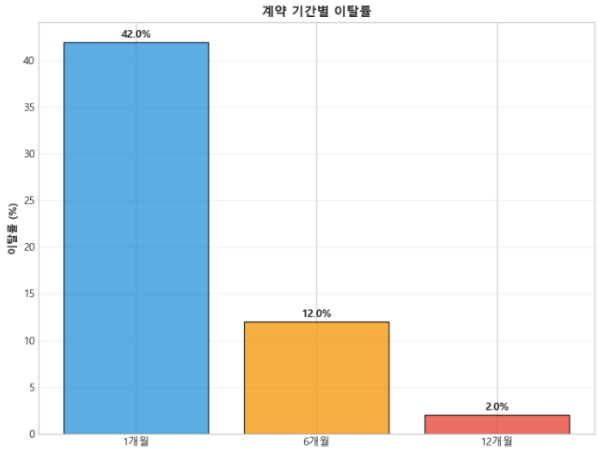
6. 주요 변수의 분포 변화 시각화

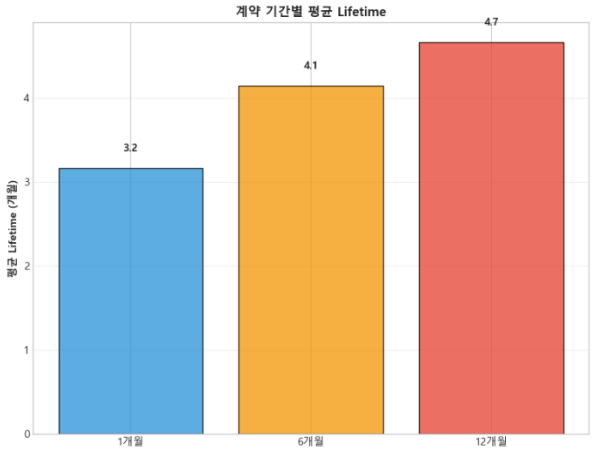
- 회원 활동 기간 (Lifetime)





- 계약 기간 (Contract_Period)





- 수업 참여 빈도별 이탈률 (Avg_class_frequency_total)