『당행의 4분기 고객 데이터 분석 보고서』

Leader 이용혁

Sub-Leader 양영화

Member 이수진 윤해민 진선영

2024.07.30

B05 - GOAT

목차

추진 배경 및 진행 결과

1. 고객데이터 분석 프로젝트 추진 배경
2. 고객데이터 분석 프로세스
3. 고객데이터 분석 결과
4. 인사이트 및 기대효과

1. 고객데이터 분석 추진 배경

1 현황 및 문제점

- 당행의 신용 등급이 낮은 고객의 비율(23%)이 타행 평균 신용 불량자 대비 10% 이상 많은 것으로 파악되며 이에 따른 관리 필요
- 마케팅 및 고객관리 차원에서 효과적인 전략 수립 및 실행을 위한 데이터 기반 근거 마련 시급

2 추진 목적

- 당행의 전반적인 고객 현황 파악
- 고객을 세분화하여 분석하고, 대출 심사 및 고위험군 고객 대응 전략 지원
- 특히, 리스크 관리가 필요한 고객 세분화 및 액션 플랜 도출 지원

3 기대 효과

- 데이터 분석 기반 고객 관계 관리 전략 개선
- 당행의 현금 유동성 리스크 감소
- 장기적으로 Data-Driven 의사결정 문화 기여

*고정이하여신(금융감독원)
고정 : 90일 이상 연체시 / 30~90일 (요주의) / 30일 미만(정상)

2. 고객데이터 분석 프로세스 – 진행 프레임

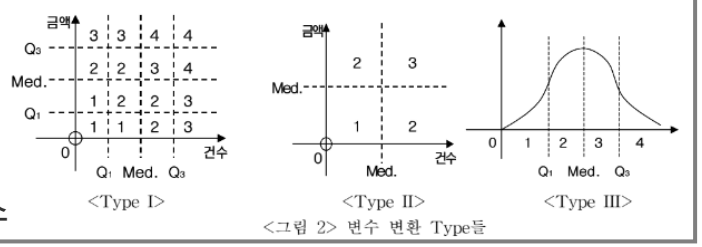
고객 분석 프로젝트는 크게 5단계를 거쳐 진행

00 은행 고객들의 연간 데이터 중 4분기(9-12月) 데이터 활용

전략방향 수립	데이터 수집	데이터 정제	데이터 분석	세그먼트 프로파일링
• 고객 데이터 분석 목표 설정 • 가설 설정 • 분석 주제 및 방향 설정 • 분석 기법 선정	• 당행의 4분기 데이터 (출처 : Kaggle) • 필요한 추가 데이터 확인	• 데이터 전처리* • 데이터 통합	• 현황 파악을 위한 데이터 EDA 및 통계 분석 • Clustering 적용 통한 고객 세분화 수행	• Segment별 특성 파악 - 프로파일 분석 • Segment 라벨링 - 프로파일 특성 파악 후 이름 부여

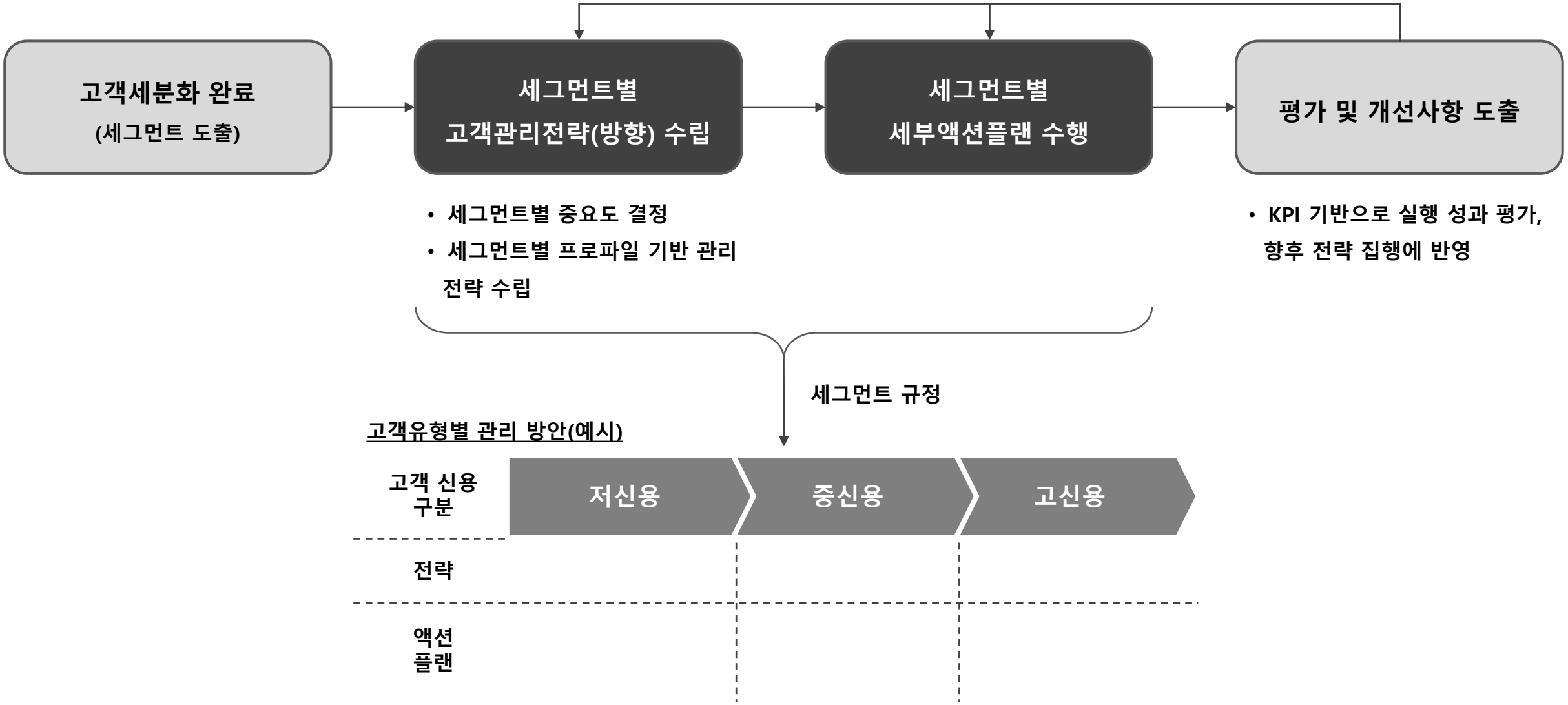
*데이터 전처리

1.결측치 처리
2.이상치 처리
3.분석목적에 맞는 새 컬럼 생성
4.분석에 필요한 변수 추출
5.분석의 효율을 위한 데이터 size 축소



<그림 2> 변수 변환 Type들

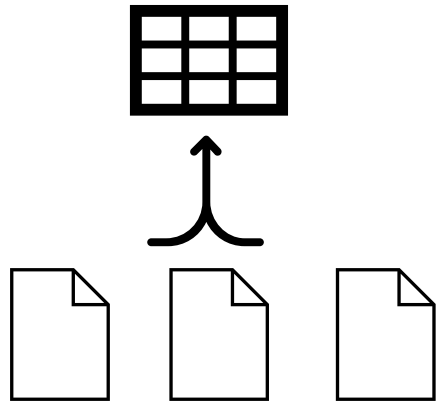
2. 고객데이터 분석 프로세스 – 이후 추진 계획

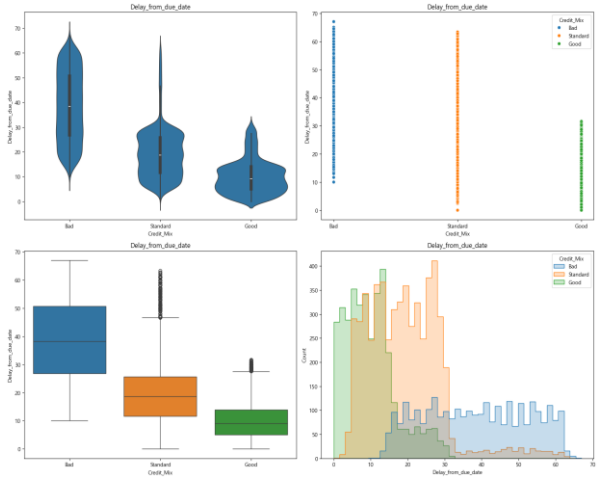


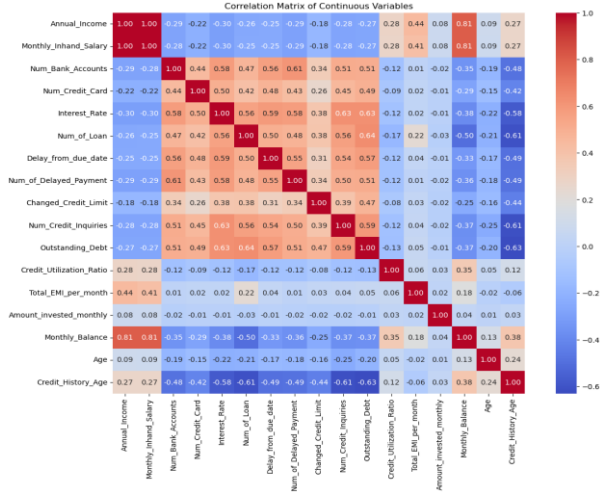
3. 고객데이터 분석 진행 개요 – 데이터 분석 및 모델링

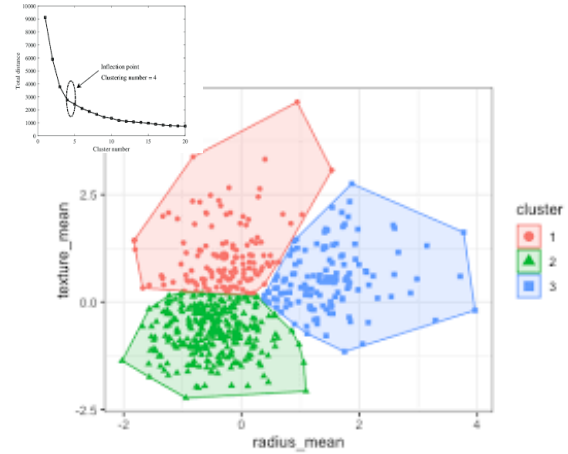
고객데이터 분석은 아래와 같은 4단계 과정으로 진행함

데이터 통합 및 전처리	EDA	통계 분석	세그먼트 분석









- 오탈자 및 결측값 처리
- 이상치 탐색 및 처리
- 고객 별 4개월 데이터 통합

- 데이터 현황을 파악하기 위한 데이터 시각화를 포함한 다양한 EDA 수행

- 변수 간 관계 파악을 위한 통계 분석

- K-means 군집화를 통한 신용 불량자 추가 세그멘테이션 진행 및 인사이트 도출

3-0. 데이터 설명 (Description)

특징	개수
데이터 개수	50,000
컬럼 수	27

컬럼	카테고리	뜻	특징	데이터타입	결측값
ID	인구통계학적 변수(6)	고유한 식별자	식별자	object	
Customer_ID		고객 식별자	식별자	object	
Name		고객 이름	식별자	object	5015
Age		고객의 나이		int	
SSN		주민등록번호	식별자	object	
Occupation		직업		object	
Annual_Income	금융 정보 변수(5)	연간 소득	고객의 연간 총 소득	float	
Monthly_Inhand_Salary		월 실수령 급여	세금 및 기타 공제를 제외한 월별 실수령 금액	float	7498
Outstanding_Debt		미결제 부채	현재까지 결제되지 않은 부채의 총액	float	
Credit_Utilization_Ratio		신용 이용 비율	사용 가능한 신용 한도 중 사용된 금액의 비율	float	
Monthly_Balance		월말 잔액	월말 기준 계좌의 잔액	float	562
Num_Bank_Accounts	거래 변수(6)	은행 계좌 수	고객이 보유한 은행 계좌의 수	int	
Num_Credit_Card		신용카드 수	고객이 보유한 신용카드의 수	int	
Total_EMI_per_month		월별 총 EMI <할부금>	고객이 매월 지불하는 EMI(원리금 균등 상환액)의 총합	float	
Amount_invested_monthly		월별 투자 금액	고객이 매월 투자하는 금액	float	2271
Num_of_Loan		대출 건수	고객이 받은 대출의 건수	int	
Type_of_Loan		대출 종류	고객이 받은 대출의 종류	int	5704
Interest_Rate	신용 변수(7)	대출 이자율	고객이 받은 대출의 이자율	float	
Delay_from_due_date		연체 기간	연체된 일수	int	
Num_of_Delayed_Payment		연체 횟수	연체된 결제의 횟수	int	3498
Changed_Credit_Limit		신용 한도 변경	신용 한도가 변경된 횟수	int	
Num_Credit_Inquiries		신용 조회 수	신용 조회가 이루어진 횟수	int	1035
Credit_Mix		신용 구성	고객의 신용 유형 구성	object	
Credit_History_Age		신용 기록 연령	고객의 신용 기록 기간	object	4470
Payment_of_Min_Amount	결제 행동 변수(2)	최소 금액 지불 여부	최소 지불 금액을 지불했는지 여부	object	
Payment_Behaviour		결제 행동	고객의 결제 행동 패턴	object	
Month		데이터가 수집된 월		object	

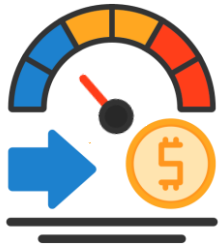
❖ **기간** : 데이터는 9월부터 12월까지의 고객별 월 단위로 구성되어 있습니다.
❖ **고객 단위** : 각 고객 당 총 4개의 데이터 포인트가 있습니다.

3-0. 데이터 설명 (Description)

- **연구 대상 : 4분기에 00은행을 이용한 12,500명**
- **집단 구분 : 신용 등급 변수를 활용하여 'Good', 'Standard', 'Bad' 로 집단을 구분**

- **집단 특성**

Good



- 총 3,701명
- 평균 연령 : **37.4세**
- 연소득 ($) : **68,516.2**
- 월급 ($) : **5,664.8**

Standard

- **총 5,646명**
- 평균 연령 : 33.4세
- 연소득 ($) : 47,630.9
- 월급 ($) : 3,943.1

Bad

- 총 2,919명
- **평균 연령 : 30.3세**
- **연소득 ($) : 32,803.7**
- **월급 ($) : 2,722.5**

3-0. 데이터 설명 (Description)

- 금융 정보 변수

	고신용	중신용	저신용	Total
	M (SD)	M (SD)	M (SD)	M (SD)
미지불 채무	740.1 (431.6)	1066.8 (686.3)	**3002.3 (1064.2)**	1428.8 (1155.5)
신용 이용률*	32.8 (3.3)	32.3 (3.0)	31.6 (3.0)	32.3 (3.1)
월간 잔액	**507.1 (225.3)**	397.6 (158.7)	281.6 (72.0)	403.0 (186.9)

- **신용 이용률** : 신용카드나 기타 신용 대출에서 사용한 신용 한도와 비교하여 실제 사용한 신용의 비율
  신용 이용률이 낮을수록 신용 점수가 높아지는 경향이 있다고 알려져 있음.

3-0. 데이터 설명 (Description)

- 거래 변수

	고신용	중신용	저신용	Total
	M (SD)	M (SD)	M (SD)	M (SD)
보유 계좌 개수	2.9 (2.0)	5.7 (1.8)	**8.0 (1.4)**	5.4 (2.6)
보유 카드 개수	4.3 (1.7)	5.3 (1.7)	**7.5 (1.7)**	5.5 (2.1)
월 할부금 총액	126.8 (198.0)	111.7 (150.9)	**141.7 (123.9)**	123.4 (161.6)
월 투자금 총액	**656.5 (1258.6)**	604.7 (1218.0)	584.2 (1287.9)	615.4 (1247.4)
대출 건수	2.0 (1.4)	3.1 (2.0)	**6.4 (1.9)**	3.5 (2.4)

3-0. 데이터 설명 (Description)

- 신용 변수

	고신용	중신용	저신용	Total
	M (SD)	M (SD)	M (SD)	M (SD)
이자율	6.4 (3.4)	14.8 (6.9)	**24.5 (5.7)**	14.5 (8.7)
마감 지연일	10.1 (6.9)	19.4 (10.1)	**38.5 (13.9)**	21.1 (14.7)
마감 지연 횟수	7.1 (4.6)	14.0 (3.8)	**20.0 (3.2)**	13.4 (6.2)
카드 사용한도 변경 횟수	6.4 (3.4)	11.4 (4.9)	**14.1 (8.9)**	10.5 (6.5)
신용 조회수	4.3 (2.1)	7.1 (3.5)	**11.4 (2.6)**	7.3 (3.9)
신용 기간	**24.6 (5.3)**	19.1 (7.4)	9.9 (5.0)	18.6 (8.3)

3-1. 데이터 통합 및 전처리 (Preprocessing)

- **결측치 및 오탈자, 이상치 처리**
  - 나이, 직업 등 **고객별 변동이 없는 변수** : 결측치 및 이상값을 각 고객의 **최빈값으로 대체**
  - 연체 이율, 투자 비용 등 **월별 변동 가능 변수** : 도메인 지식 및 IQR 기반 이상치 처리, 처리 불가한 경우는 탈락
- **그룹화 및 집계**
  - 고객별 4개월의 데이터를 하나의 행으로 통합
  - 통합 시 분석에 필요한 변수 선택
  - 고객 ID(Customer_ID 변수)를 기준으로 데이터 그룹화
    - **최빈값** : Occupation, Annual_Income, Monthly_Inhand_Salary 등
    - **평균** : Num_Bank_Accounts, Interest_Rate
    - **최대값(최신 데이터)** : Age, Credit_History_Age 등

최종 데이터셋

- 12,500명의 고객 중 **12,265명의 고객 데이터 유지**
- 통합된 데이터프레임 생성

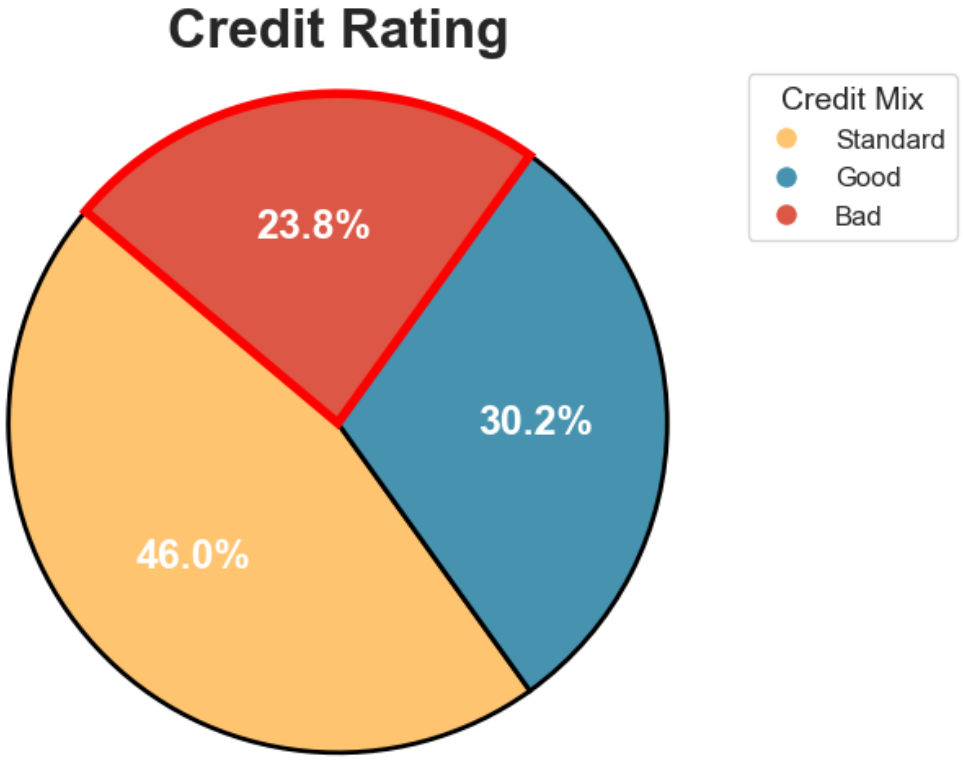
ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income
0x160a	CUS_0xd40	September	Aaron Maashoh	23	821-00-0265	Scientist	19114.12
0x160b	CUS_0xd40	October	Aaron Maashoh	24	821-00-0265	Scientist	19114.12
0x160c	CUS_0xd40	November	Aaron Maashoh	24	821-00-0265	Scientist	19114.12
0x160d	CUS_0xd40	December	Aaron Maashoh	24_	821-00-0265	Scientist	19114.12

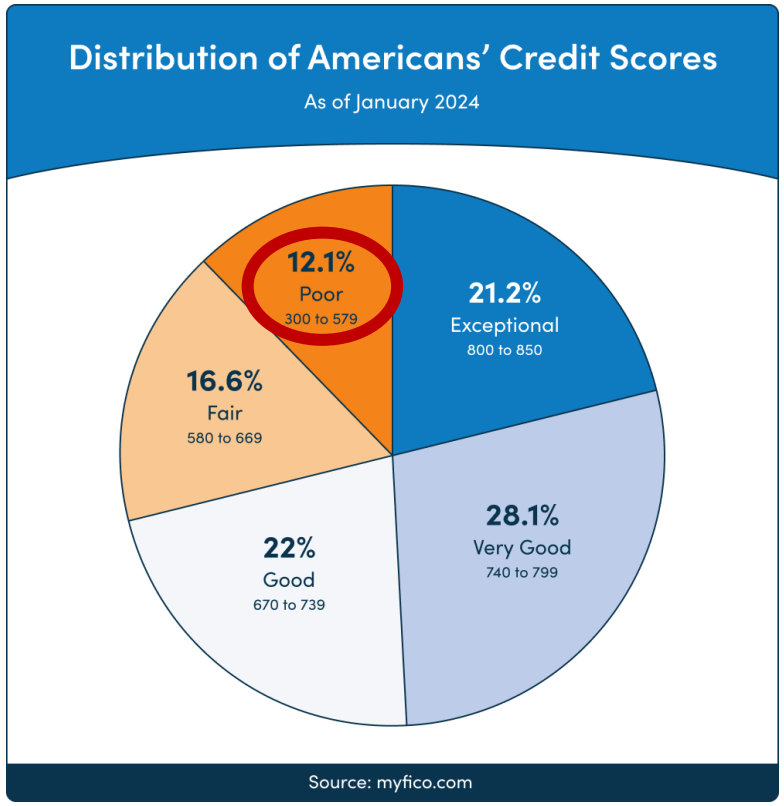
Customer_ID	Occupation	Annual_Income	Age	…
CUS_0xd40	Scientist	19114.12	24	…

Q0. 당행의 문제점은?

3-(2,3). EDA 및 통계 분석

신용 등급별 비율





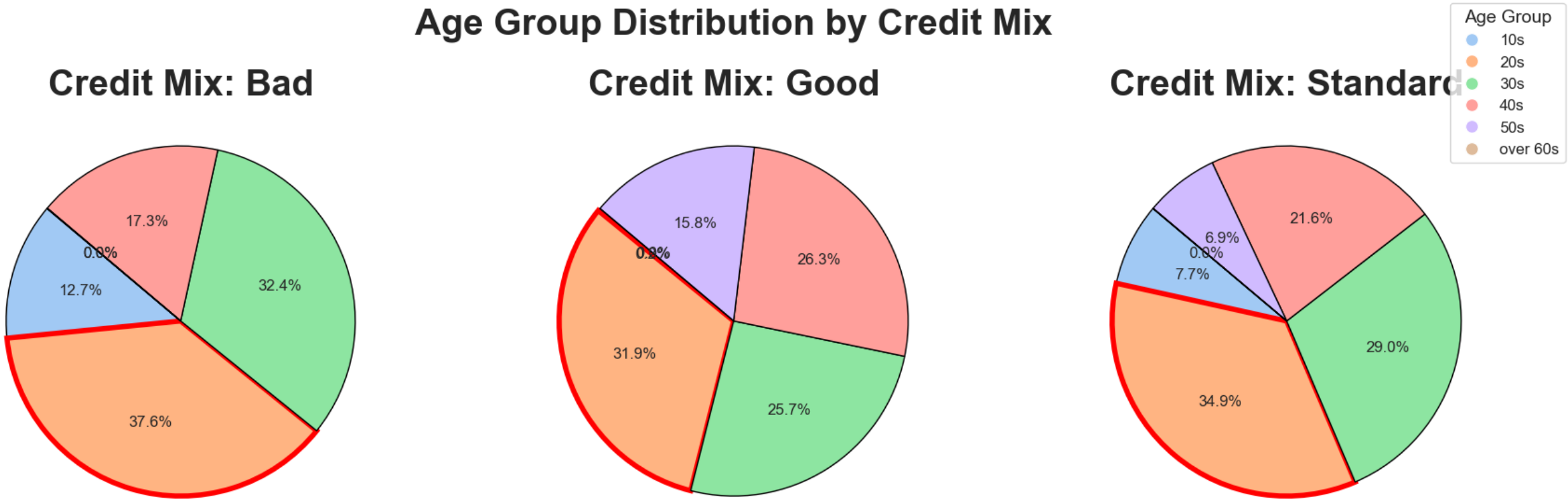
https://upgradedpoints.com/credit-cards/credit-score-facts-statistics/

당행의 **저신용 고객**은 전체 고객의 **23.8%** 이며, 2024년 1월 기준 미국 타행 저신용 고객 평균 비중 **12.1%** 보다 **11.7%** 많은 상황

Q1. 당행의 고객은 연령대별로 어떠한 신용 분포를 가지고 있을까?

3-(2,3). EDA 및 통계 분석

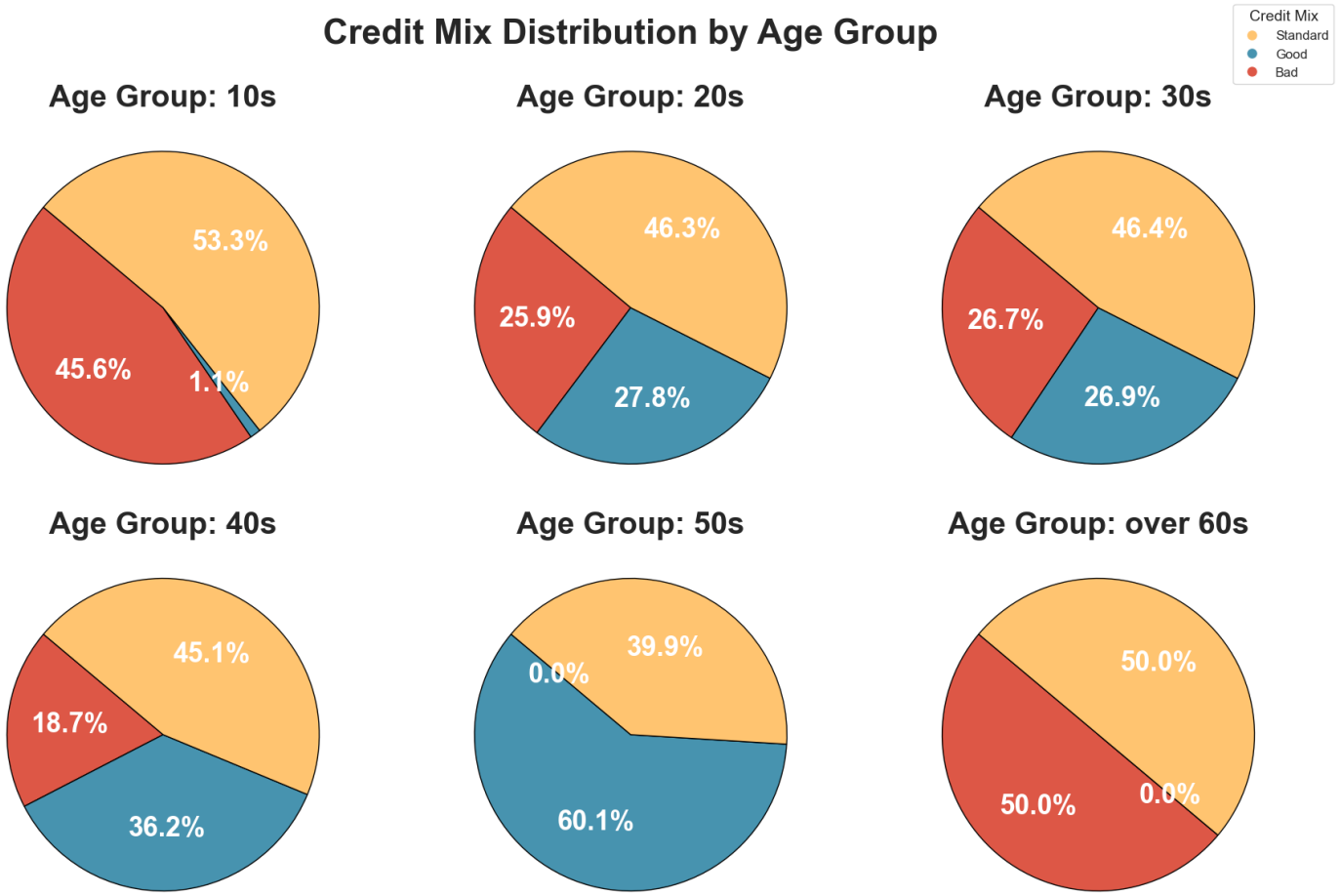
신용 상태 별 연령대 분포



- Bad, Good, Standard 모두 **20~30대 비중이 큼**
- **신용 불량의 10% 이상이 10대**인 것이 주목할 점 => 추가적인 데이터 수집 및 특성을 파악해서 왜 10대에 신용 불량이 많은 지 확인할 필요가 있음

3-(2,3). EDA 및 통계 분석

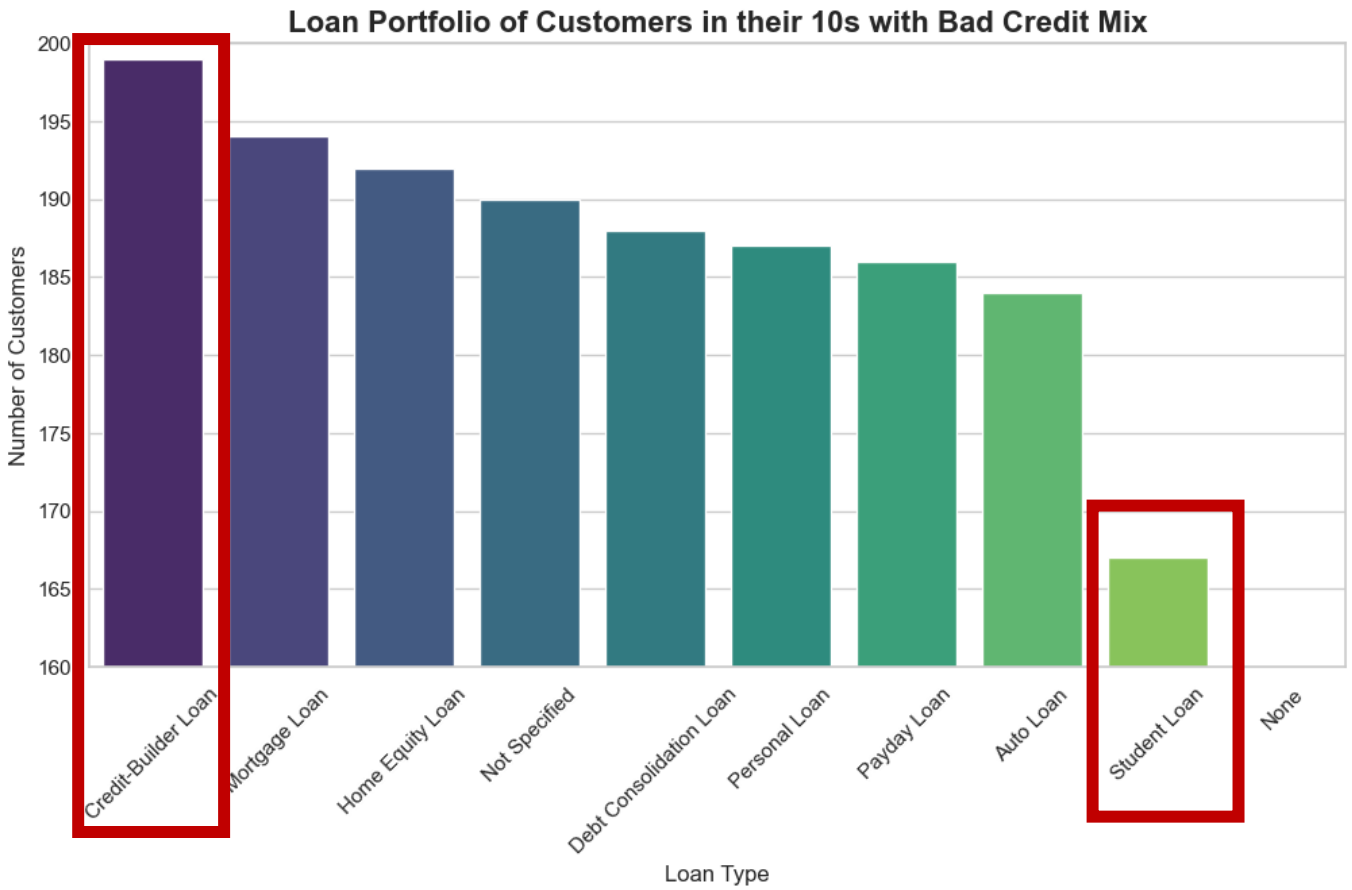
연령대별 신용 상태 분포

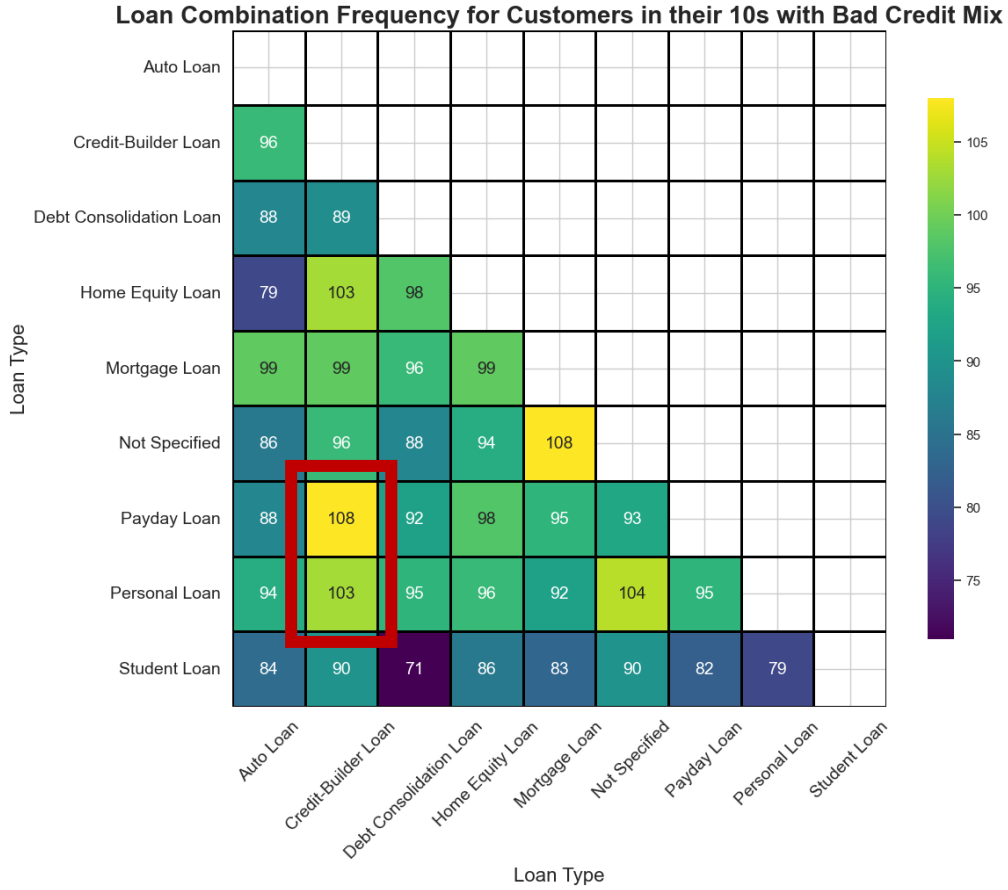


- 10 대 : Good 보다 **Bad** 의 비율이 **45.6%** 로 많다.
- 20, 30대 : Good과 Bad 의 비율이 고르게 분포. **ex) 25.9%, 27.8% / 26.7%, 26.9%**
- 40대, 50대 : Good 비중이 높다. **ex) 36.2%, 60.1%**

3-(2,3). EDA 및 통계 분석

10대 저신용 고객 대출 포트폴리오





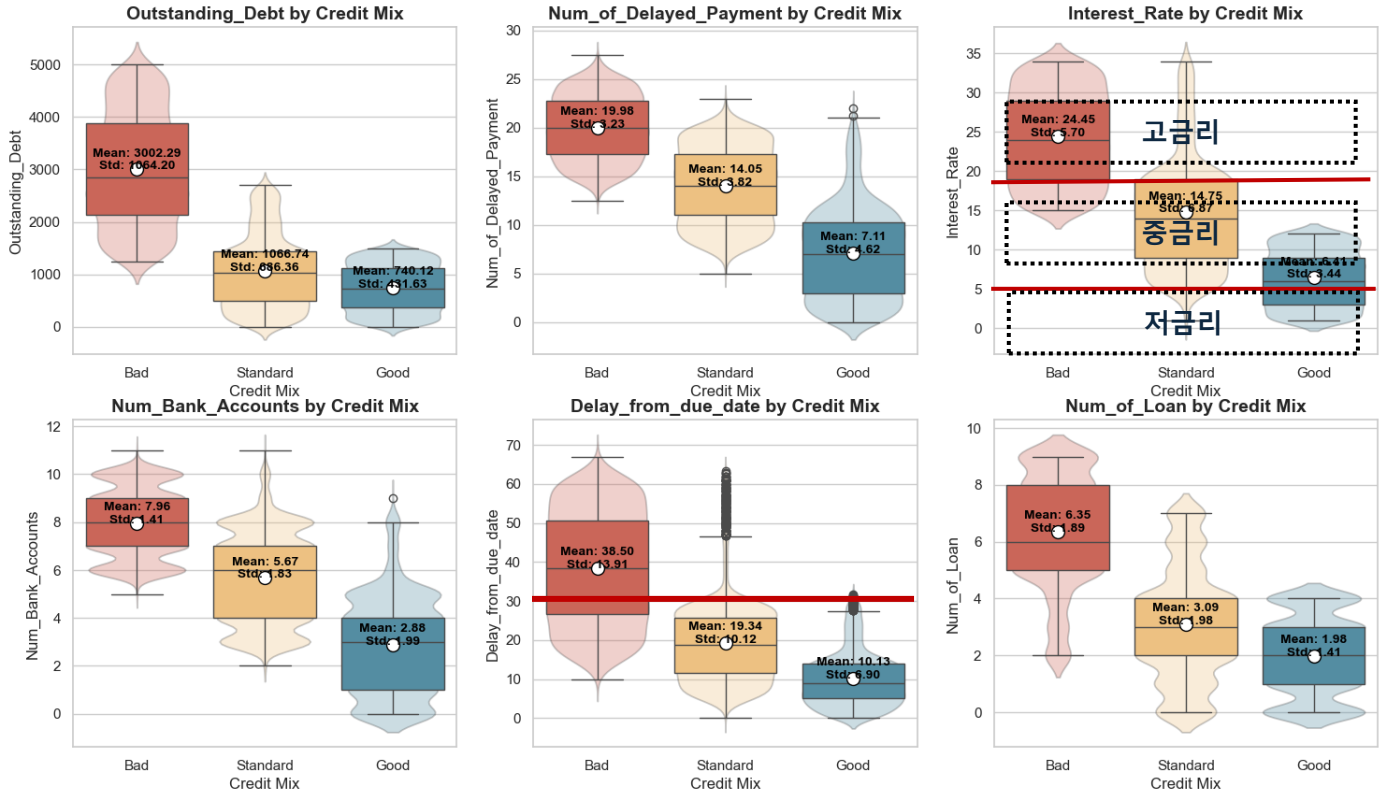
- 10대의 경우 Student Loan 이 많을 것으로 가정했으나, Credit-Builder Loan 이 많은 것이 특징이다.
- **Credit-Builder Loan** 은 **Payday Loan** 및 **Personal Loan** 과 가장 조합을 많이 이루는 것을 확인할 수 있었다.
- 이러한 점을 미루어 보았을 때, 학생 개인의 대출 보다는 **자녀 명의 대출 의심 가능 -> 추가 조사 필요**

Q2. 당행의 신용 등급별로 그룹을 나눴을 때 가장 뚜렷한 차이를 보이는 변수들은 어떤 것들이 있을까? 해당 변수들 중 특이한 특성이 있을까?

3-(2,3). EDA 및 통계 분석

신용 카테고리 별 연속형 변수의 ANOVA 분석 결과.

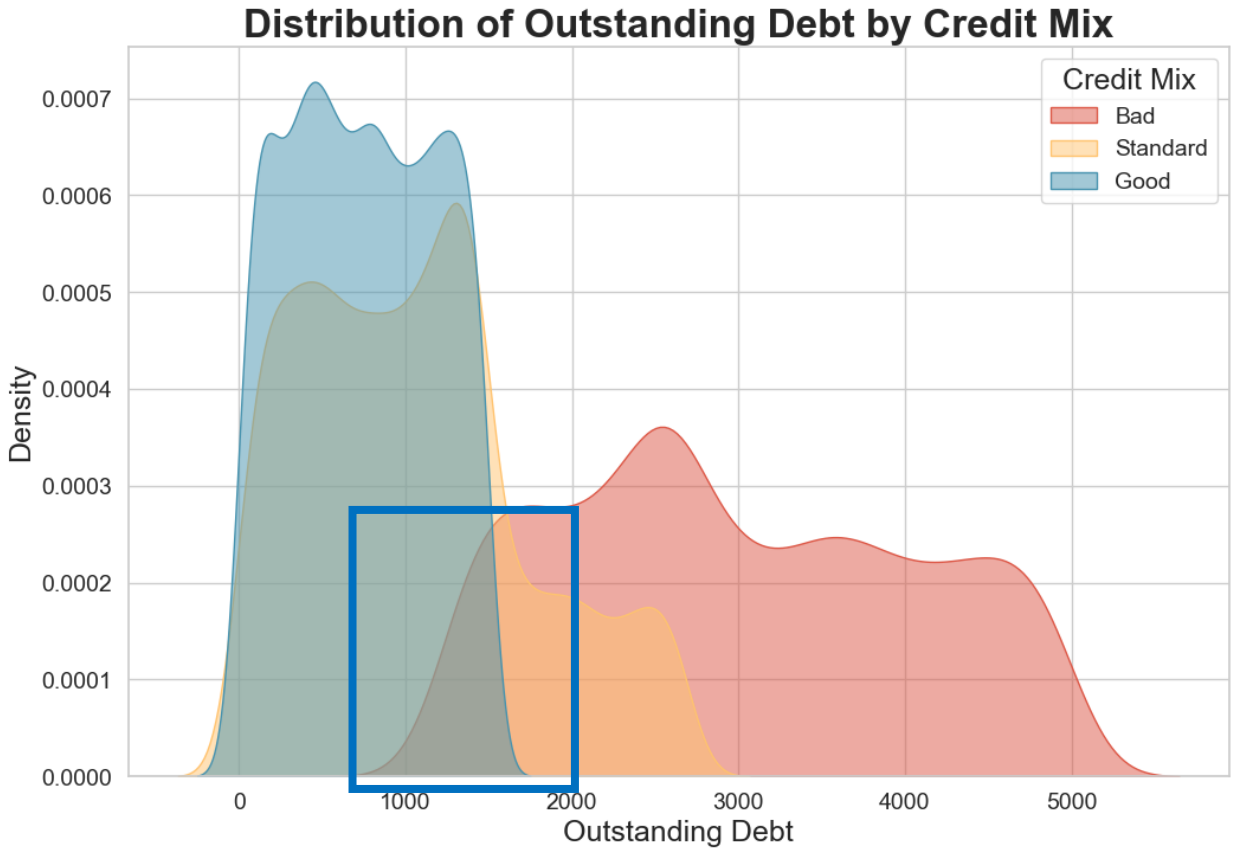
Variable	F-Value	P-Value	Effect Size (Eta Squared)
Outstanding_Debt	8959.78	0	0.999888421
Num_of_Delayed_Payment	8787.13	0	0.999886229
Interest_Rate	8055.68	0	0.9998759
Num_Bank_Accounts	6737.44	0	0.999851622
Delay_from_due_date	6263.50	0	0.999840397
Num_of_Loan	5083.36	0	0.999803351
Num_Credit_Inquiries	4857.99	0	0.99979423
Num_Credit_Card	2808.88	0	0.999644172
Changed_Credit_Limit	1567.67	0	0.999362625
Annual_Income	838.97	0	0.998809689
Monthly_Inhand_Salary	828.18	0	0.998794188
Credit_Utilization_Ratio	127.62	1.39E-55	0.992226589
Total_EMI_per_month	34.51	1.13E-15	0.97184456



- 신용 등급별로 상위 5개 변수선정. **(미결제 부채의 총액 > 연체된 결제의 횟수 > 대출 이자율 > 은행 계좌 수 > 연체된 일수)**
- 고신용 고객과 저신용 고객의 차이가 명확하게 나타나는 것을 확인.

3-2-1. Outstanding Debt (미결제 부채)

신용 등급별 Top 5 특징 변수 추가 분석



Statistics

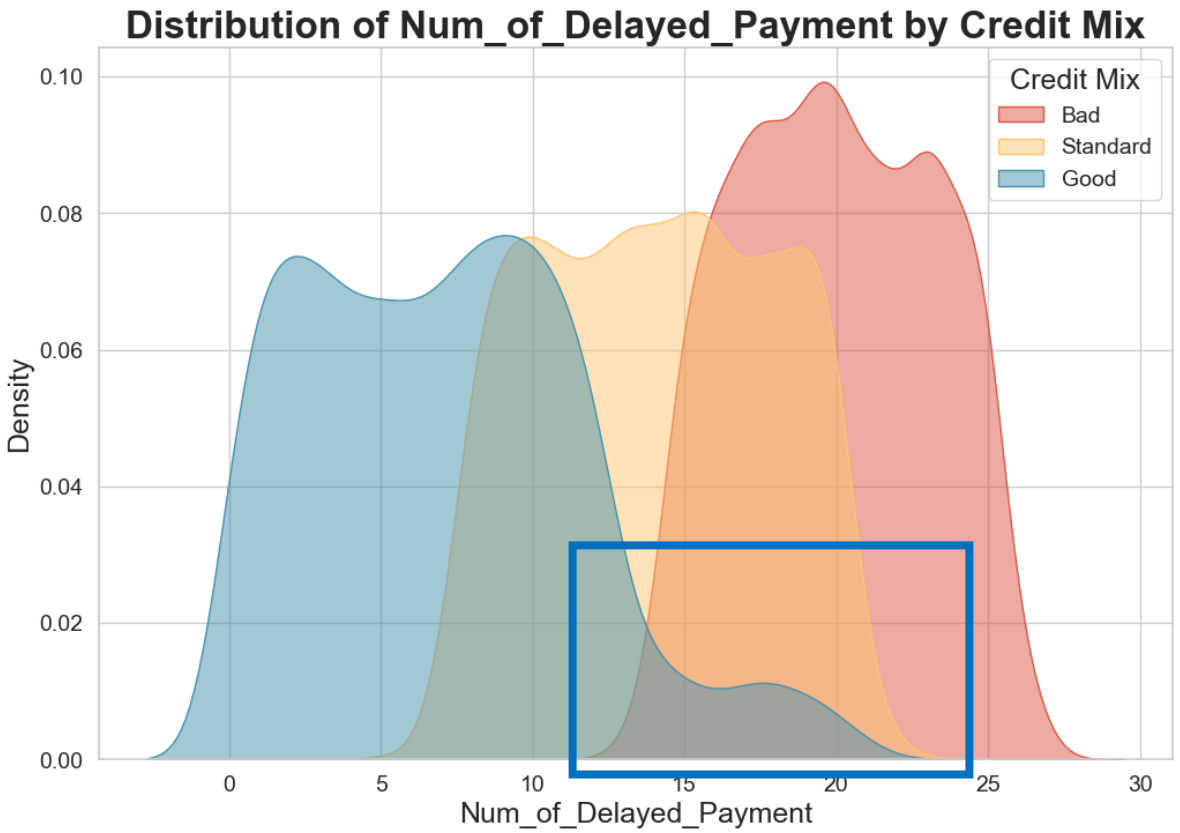
	GOOD	BAD
범위	0 to 1,500($)	1,000 to 5,000($)
분포 형태	밀도가 비교적 높음 (정규 분포)	밀도가 비교적 낮음 (정규 분포)
평균	740($)	3,002($)
편차	431.6($)	1,064.2($)
최댓값	1,498($)	4,998($)
최솟값	0.23($)	1,250($)

가설 : 신용도 등급 간에 수치적으로 중첩되는 부분이 없을 것이다.

검정 결과 : 1000~2000($) 값이 집단별로 중첩되는 것을 확인했다.

3-2-2. Delayed Payment (마감 지연 건수)

신용 등급별 Top 5 특징 변수 추가 분석



Statistics

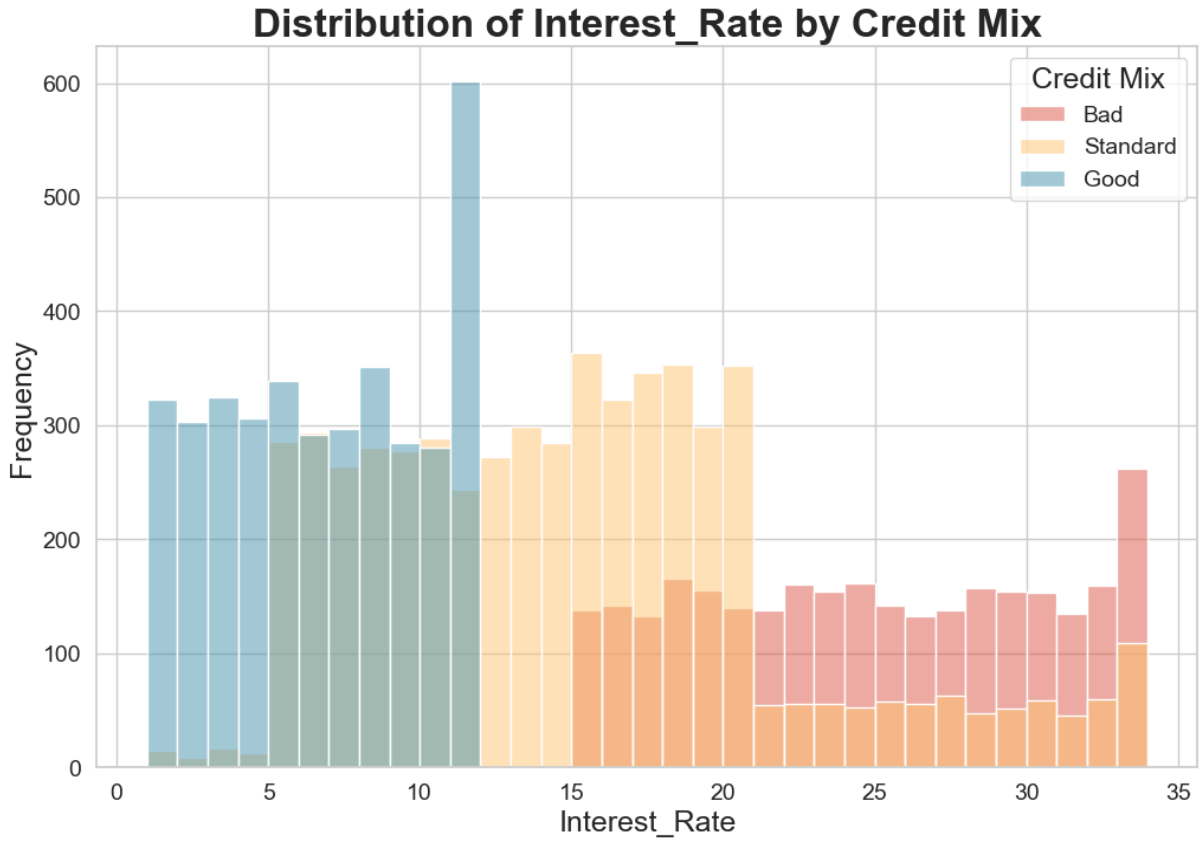
	GOOD	BAD
범위	0 to 22(회)	12 to 28(회)
분포 형태	오른쪽으로 긴 꼬리 형태	밀도가 높은 정규 분포
평균	7.1(회)	20(회)
편차	4.6(회)	3.2(회)
최댓값	22(회)	27.5(회)
최솟값	0(회) *지연 횟수가 없는 경우가 포함*	12.5(회) *항상 지연 발생*

가설 : 신용도 등급 간에 수치적으로 중첩되는 부분이 없을 것이다.

검정 결과 : 12~22(회) 값이 집단별로 중첩되는 것을 확인했다.

3-2-3. Interest Rate (이자율)

신용 등급별 Top 5 특징 변수 추가 분석



Statistics

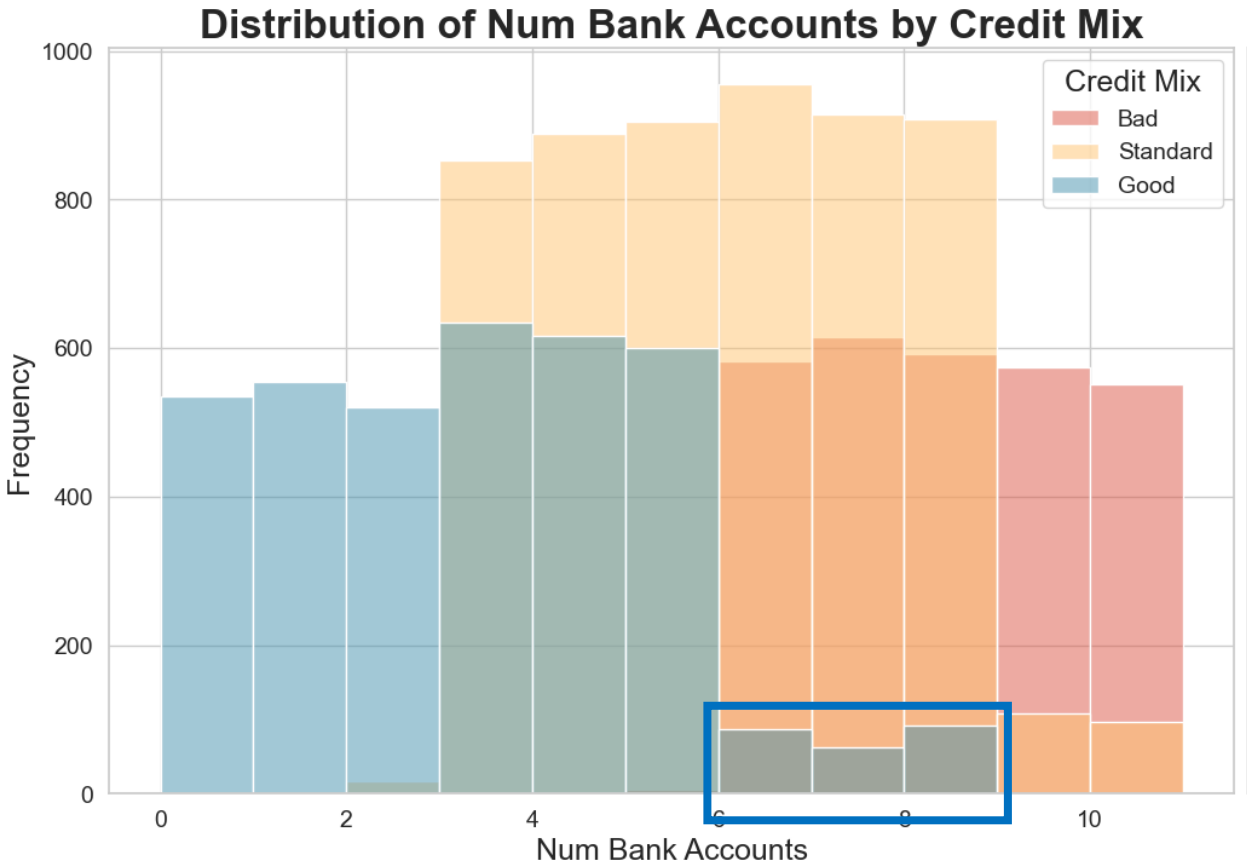
	GOOD	BAD
범위	1 to 12(%)	15 to 35(%)
분포 형태	밀도가 높은 정규 분포	밀도가 낮은 정규 분포
평균	6.41(%)	24.45(%)
편차	3.44(%)	5.70(%)
최댓값	12.0(%)	34.0(%)
최솟값	1.0(%)	15.0(%)

가설 : 신용도 등급 간에 수치적으로 중첩되는 부분이 없을 것이다.

검정 결과 : GOOD 최댓값 12% ~ BAD 최솟값 15%로 가설이 옳다.

3-2-4. Num Bank Account (보유 은행 계좌 개수)

신용 등급별 Top 5 특징 변수 추가 분석



Statistics

	GOOD	BAD
범위	0 to 9(개)	5 to 11(개)
분포 형태	유사 균등분포	유사 균등분포
평균	2.88(개)	약 2.7배 → 7.96(개)
편차	1.99(개)	1.71(개)
최댓값	8.0(개)	약 1.4배 → 11.0(개)
최솟값	0.0(개)	5.0(개)

가설 : 신용도 등급 간에 수치적으로 중첩되는 부분이 없을 것이다.

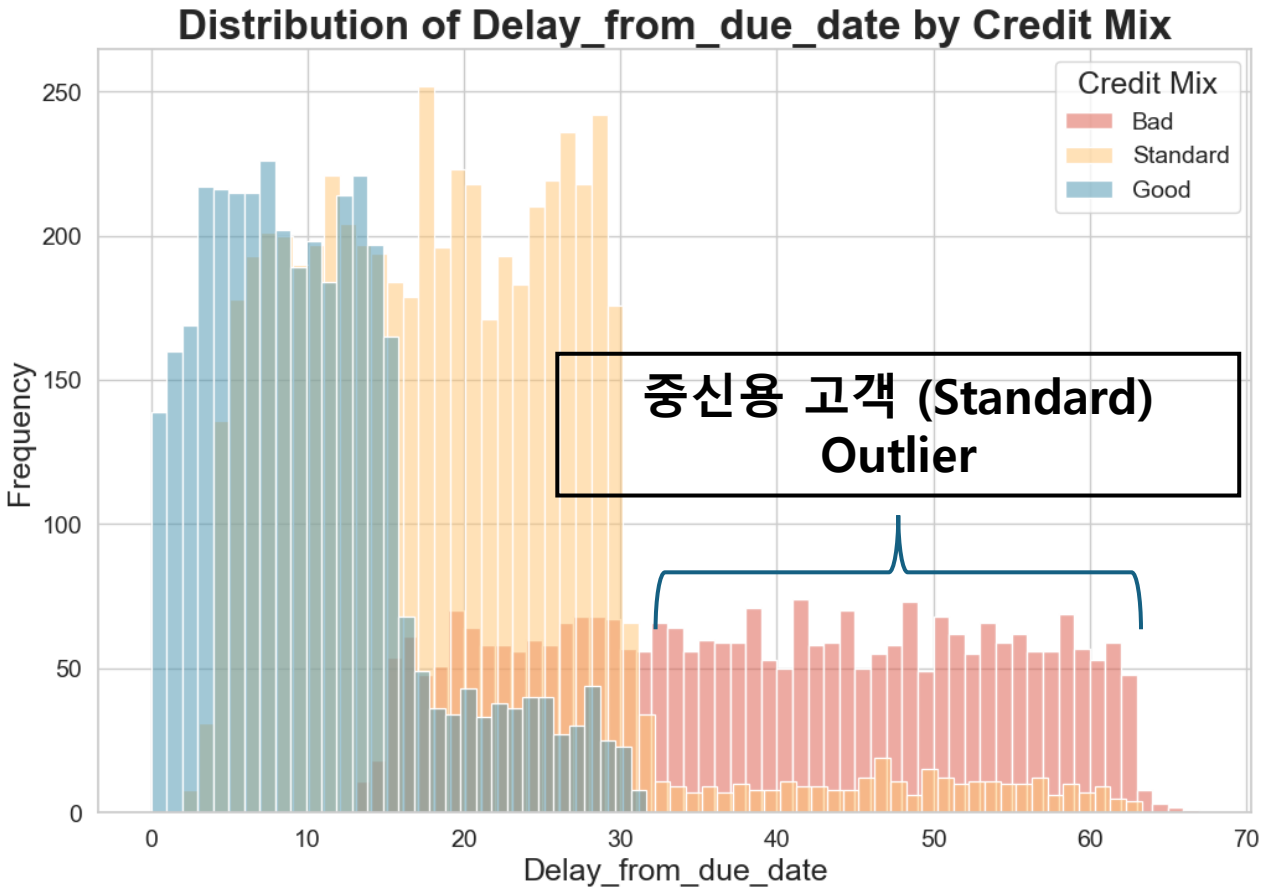
검정 결과 : 5 ~ 9개 값이 집단별로 중첩되는 것으로 확인했다.

*특이한 점 : 고신용 고객의 경우에도 은행 계좌 보유수가 높은 경우가 있음. But, 저신용 고객의 경우 은행 계좌 보유수가 **최소 5개 이상**인점은 주목할 점

저신용 고객과 은행 계좌 보유수 간의 관계 파악 분석 필요

3-2-5. Delay from due date (마감 지연일)

신용 등급별 Top 5 특징 변수 추가 분석



Statistics

	GOOD	BAD
범위	0 to 31 (일)	10 to 67 (일)
분포 형태	오른쪽으로 꼬리가 긴 분포	균등분포
평균	약 10(일)	약 38.5(일) (약 3.8배)
편차	6.9(일) *편차의 차이가 작다.*	13.9(일) (약 2 배) *편차의 차이가 크다.*
최댓값	31(일)	67(일) (약 2.1배)
최솟값	0(일)	10(일)

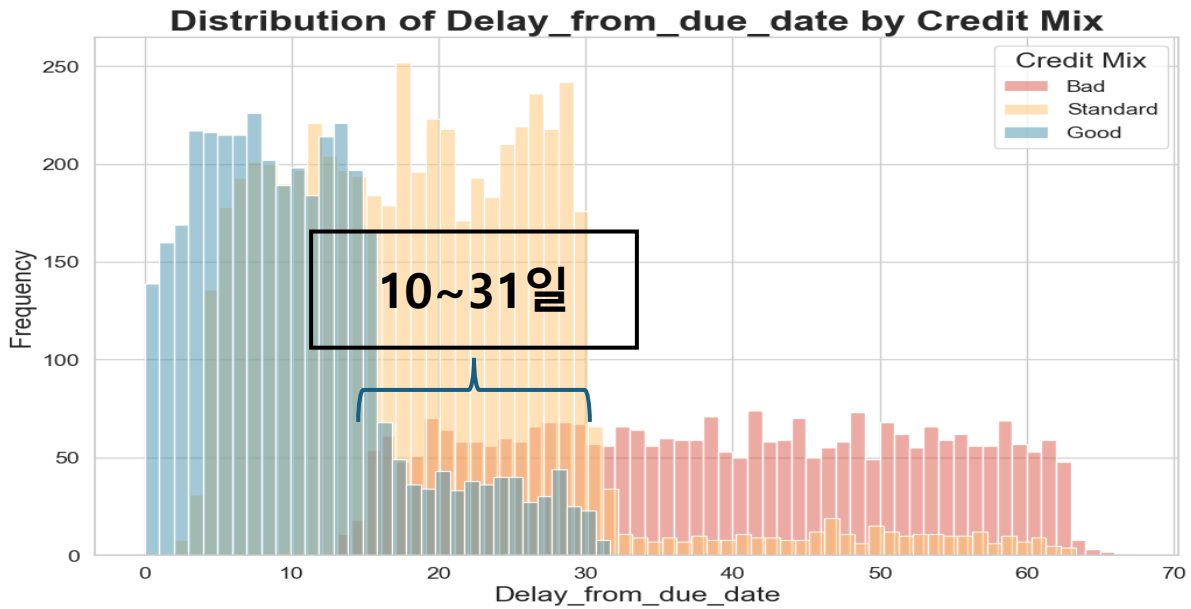
가설 : 신용도 등급 간에 수치적으로 중첩되는 부분이 없을 것이다.

검정 결과 : 10일 ~ 31일에 속한 값이 집단별로 중첩되는 것으로 확인했다.

3-2-6. 전체 데이터 기초 통계량과의 비교

신용 등급별 Top 5 특징 변수 추가 분석

Variable	F-Value	P-Value	Effect Size (Eta Squared)
Outstanding_Debt	8959.78	0	0.999888421
Num_of_Delayed_Payment	8787.13	0	0.999886229
Interest_Rate	8055.68	0	0.9998759
Num_Bank_Accounts	6737.44	0	0.999851622
Delay_from_due_date	6263.50	0	0.999840397

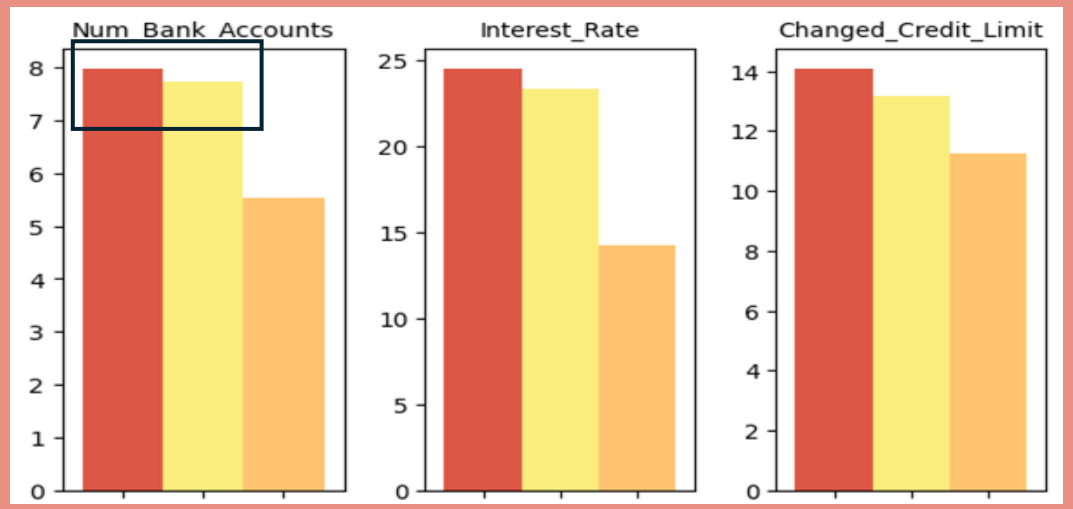


전체 데이터		Bad	Good	중첩 데이터 (10~31일)		Bad	Good
미지불 채무	mean	3002.3	740.1	미지불 채무	mean	3014	747
	std	1064.2	431.6		std	1060	432
마감 지연 건수	mean	19.98	7.11	마감 지연 건수	mean	19.9	7.9
	std	3.23	4.62		std	3.2	4.9
이자율	mean	24.45	6.41	이자율	mean	24.3	6.5
	std	5.7	3.44		std	5.6	3.4
보유 계좌 개수	mean	7.96	2.88	보유 계좌 개수	mean	8	3.2
	std	1.41	1.99		std	1.4	2.1

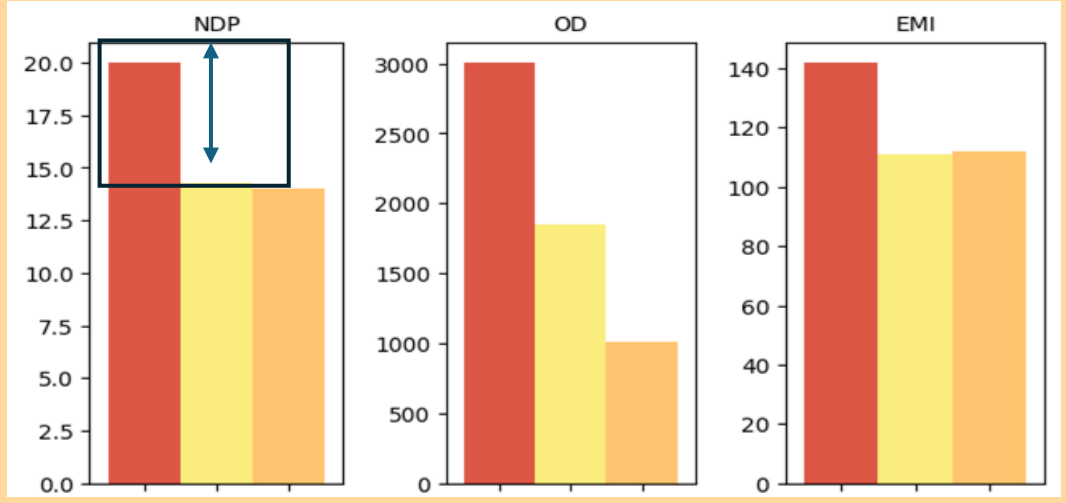
3-2-5-1. 중신용 고객의 이상치에 대한 추가 분석

저신용 고객 VS 중신용 Outlier(>=31) 그룹 VS 중신용 非 Outlier 그룹 비교

[similar variables]



[different variables]

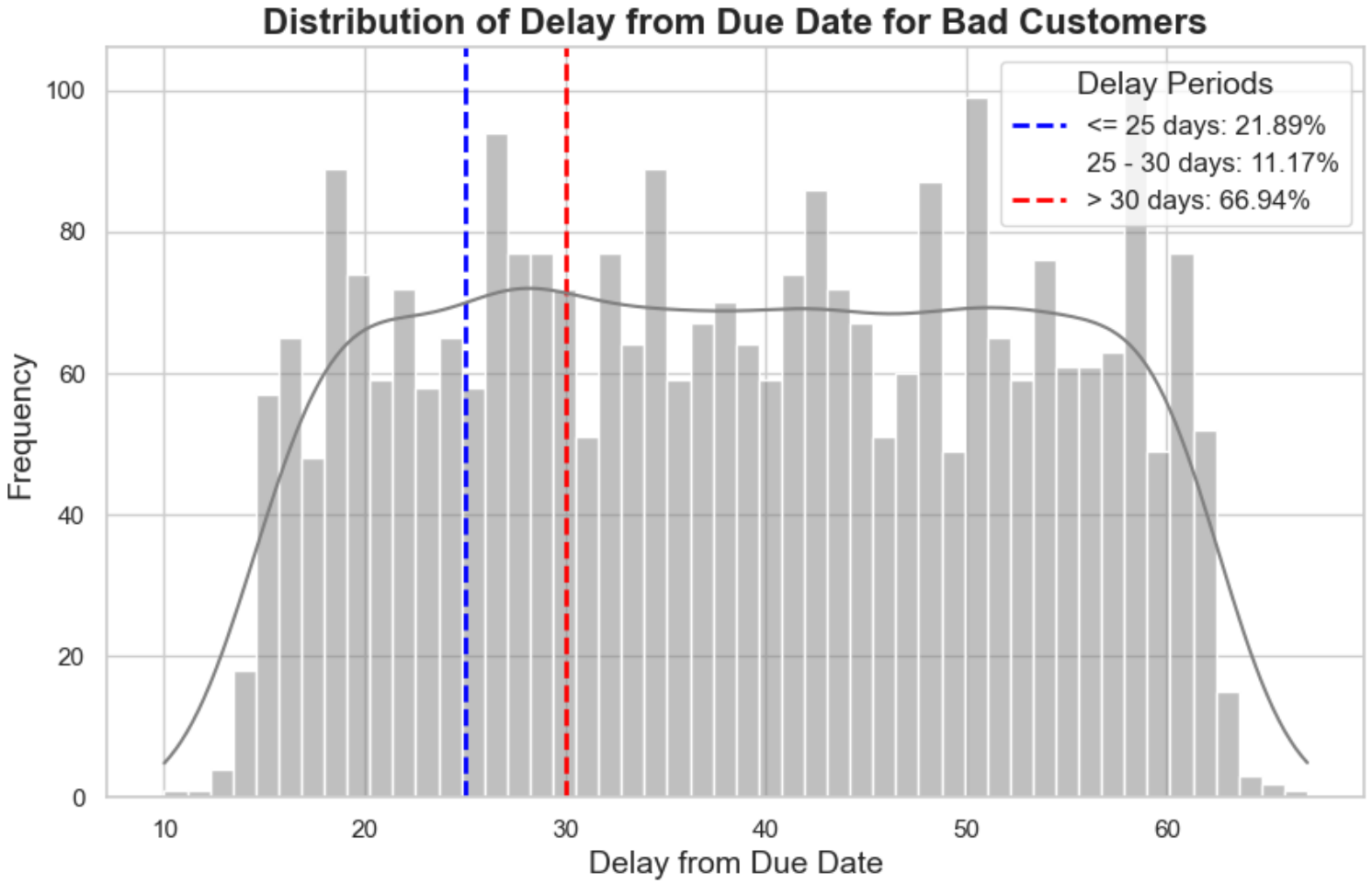


- **중신용(Standard) Outlier 그룹**과 **저신용(Bad) 고객 그룹**의 변수 평균 비교 **(유사)**
  - **ex) 은행 계좌 수, 이자율, 신용 한도 변경 횟수**
  - 중신용 그룹에서 **연체 지연일이 31일 이상 되는 고객들은 저신용 고객들과 부분적으로 유사한 경향성을 가짐.**
- **중신용(Standard) Outlier 그룹**과 **저신용(Bad) 고객 그룹**의 변수 평균 비교 **(확연한 차이)**
  - **ex) 연체 횟수, 미지불 채무, 월 할부금**
  - 저신용 고객 그룹과 유사한 변수가 다수 존재하지만, **중신용 Outlier 고객의 신용도가 중신용 고객에 속하는 원인**은 신용도에 결정적인 영향을 미치는 요인이 저신용 고객 그룹과 확연히 차이가 나는 변수일 수 있고, 다양한 변수의 조합으로 신용도를 측정하기 때문일 것으로 추정할 수 있습니다.
- **결론적**으로 신용도를 산출할 때는 다양한 변수를 고려하며, 그 중에서도 가중치가 높은 변수들이 있는 것으로 추론. 가중치가 높을 것으로 판단되는 변수들에 대한 관리가 필요.

Q3. 당행의 저신용 고객들 중에서 집중 관리가 필요한 고객들은 어떤 고객들일까?

3-(2,3). EDA 및 통계 분석

저신용 고객의 연체 일수 분포 확인



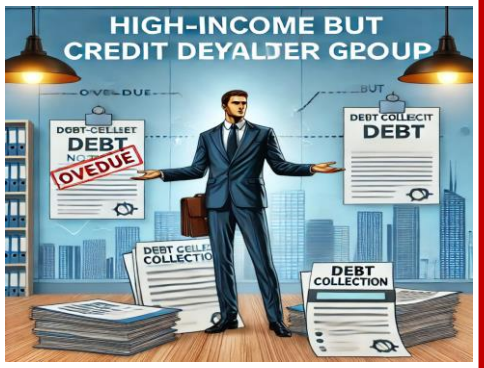
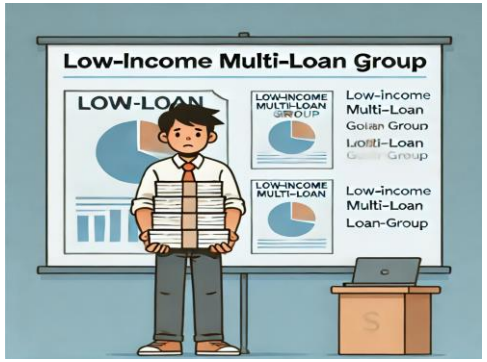
단계	연체기간
정상(Normal)	1개월 미만
요주의 (Precautionary)	3개월 미만
고정(Substandard)	3개월 이상
회수의문(Doubtful)	3개월이상 ~ 1년미만 대출자나 대출처의 채무사오한 능력이 현저하게 악화되어 채권회수에 심각한 위험이 발생한 대출금
추정손실 (Estimanted loss)	1년 이상

금감원 기준 요주의 고객(30일 이상) 약 67% -> 모니터링 및 대응 방안 강구
요주의 고객 전 단계의 고객 (25~30일) 요주의 고객 전환 방지

3-4. 세그먼트 분석

저신용 고객의 세부 세그먼트 분석을 위해 PCA + K-means Clustering을 활용해 4개의 클러스터 생성

집중 관리 대상

고소득군 (저신용 고객 대비 24.3%, 전체 대비 5.7%)	최저소득 고위험 투자군 (저신용 고객 대비 24.3%, 전체 대비 5.7%)	고소득 다중대출군 (저신용 고객 대비 15.6%, 전체 대비 3.7%)	저소득 다중대출군 (저신용 고객 대비 35.6%, 전체 대비 8.4%)
			
연간 소득: 53,528(높은 소득 수준)	**연간 소득: 16,672 (가장 낮은 소득 수준)**	**연간 소득**: 59,630 (가장 높은 소득 수준)	**연간 소득**: 19,310
월 소득: 4,459	**월 소득: 1,388**	**월 소득**: 4,932	**월 소득**: 1,601
대출 수: 4.83	**대출 수: 4.81 (가장 적은 대출 수)**	대출 수: 7.46 (가장 많은 대출 수)	**대출 수**: 7.15
미지급 금액: 2,201	**미지급 금액: 2,032**	**미지급 금액**: 3,575 (가장 높은 미지급 금액)	**미지급 금액**: 3,500
월 잔액: 366	**월 소득 대비 투자 비율: 40% (가장 높은 투자 비율)**	**월 EMI**: 311 (가장 높은 EMI)	**월 EMI**: 91
신용 이력 기간: 12.78년	**신용 이력 기간: 13.41년 (가장 긴 신용 이력)**	**신용 이력 기간**: 7.81년 (가장 짧은 신용 이력)	**신용 이력 기간**: 8.07년
기타 특징: 높은 월 소득과 잔액, 안정적인 대출 관리	**기타 특징: 낮은 소득과 높은 투자 비율, 높은 위험성**	**기타 특징**: 높은 소득과 대출 수, 높은 미지급 금액	**기타 특징**: 낮은 소득과 다수의 대출, 높은 미지급 금액

3-4. 요약 (인사이트 및 기대효과)

대출 심사 및 현금유동성 리스크를 줄이기 위한 CRM 우선 순위 고객

- 중’저 신용 고객의 연체 일수 : 25~30일 (2.6%)
- 중’저 신용 고객의 연체 일수 : 30일 이상 (16%)
- 저신용 고객 중 최저신용 고위험 투자군(전체 중 5.7%)

추가 분석 수행 대상

- 10대 저신용 고객 (대출 포트폴리오에서 특이점 확인)
- 저신용 고객의 다중 은행 계좌수를 가지는 원인
- 중신용 연체 outlier 고객 그룹 분석을 통해 확인한 변수 중요도 인사이트를 ML 모델의 변수 중요도와 비교하여 유효성 검증
- 대출심사 및 개인 맞춤형 CRM 전략을 위해 중,고 신용도 고객 그룹에도 세분화된 고객 분석 추진

기대효과

- 데이터 분석 기반 CRM 전략 개선
- 당행의 현금 유동성 리스크 감소
- 장기적으로 Data-Driven 의사결정 문화 기여

이용혁(Data_3기), 양영화(Data_3기), 윤해민(Data_3기), 진선영(Data_3기), 이수진(Data_3기)님이 입장하였습니다.

짧은 시간이지만 하나의 목표와 성취를 이루기 위해 좋은 팀원들과 함께 호흡을 맞출 수 있어 재미있었고, 많은 회의와 고민을 하며 생각의 깊이가 한 단계 깊어진 것 같아 너무 의미 있는 시간이었습니다.

팀원 분들 덕에 프로젝트 기간동안 정말 많은 걸 배웠어요! 이번에 해소되지 않은 의문점을 해결할 방법을 앞으로 배우게 될 거라 생각하니 기대가 되기도 합니다... 1주일동안 다들 너무 고생하셨고 감사했습니다!!

조원분들 덕분에 많은 것을 배울 수 있었고, 새로운 경험을 할 수 있었습니다. 기초를 다지는 것도 쉽지 않았는데, 심화 과정에서는 어떤 새로운 기술을 배우고 이를 어떻게 응용할 수 있을지 궁금하기도 하고 기대되기도 합니다.

이번 팀프로젝트를 하며 부족한 점이 많다는 것을 깨달았지만 옆에서 팀원 분들이 많이 도와주셔서 해낼 수 있고, 이로 인해 배운 것들이 많아 뜻 깊고 알찬 시간이었습니다.

많이 부족하다고 느끼고 있었는데, 그 부족함을 절실하게 깨닫게 해줬던 프...로....젝....트.....(맵다 매워) 앞으로 열심히 하겠습니다. 모두들 고생 많으셨고, 함께 해주셔서 감사했습니다!!!!

(소감이라....그냥 힘들어요 모르겠고 힘들어요.....)

전송

감사합니다.

End of Document

Q&A

1) ANOVA 분석을 사용한 이유

Q&A

1. One-way ANOVA 란 무엇일까요?

- 세 집단 이상의 집단을 비교할 때 사용하는 분석 방법
- 독립변수와 종속변수가 1개
- 영가설과 연구 가설
  - 영가설 : 세 집단 간 차이가 없다.
  - 연구 가설 : 적어도 한 집단에서 차이가 있다.
- 사후 분석 : 실제 어떤 집단에서 차이가 있었는지 확인하기 위해 실시

2. 사용한 이유

• 본 연구에서는 신용도에 따라 집단을 구분하여 분석을 실시했습니다. 신용도는 총 3개로 구분되어 있어 3개의 집단을 비교할 수 있는 one-way ANOVA를 사용했습니다.

2) PCA를 사용한 이유

PCA(Principal Component Analysis) 란 데이터 정보를 압축하는 기능이 있습니다. 따라서, 우리의 데이터에서 다양한 변수를 고려하고 싶었고, 이에 따라 주성분 2가지를 뽑아 k means clustering 을 수행하기 위해 사용했습니다.

PCA의 개념

주성분 분석*이라고 불리며, 고차원의 데이터를 저차원의 데이터로 환원시키는 기법.* ***표본의 차이를 가장 잘 나타내는 '주성분'을 선정한다.*** *주성분이란, 전체 데이터의 분산을 가장 잘 설명하는 성분. 즉, 효과가 좋은 X값들을 솎아내는 작업이라고 볼 수 있다.*

scikit-learn 패키지에서 제공.

3) Clustering 기법으로 K means 를 선택한 이유.

각 군집의 평균을 기준으로 거리를 계산하여 클러스터링을 해주는 Kmeans 군집 방법이 가장 이해하기 쉬웠고, 복잡한 기법을 사용하지 않아도 군집을 적절하게 잘 구분하는 것을 시각화로 확인했습니다. 또한, 계산 비용 효율적인 측면과 모델 설명력을 기반으로 kmeans 를 사용했습니다.

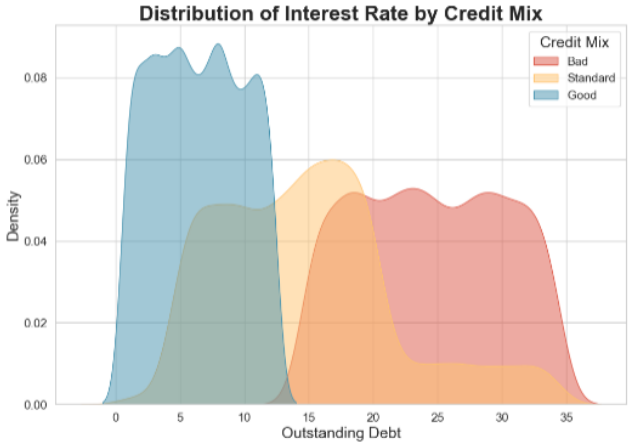
Clustering 개념

*군집화, 클러스터링. **비지도학습**의 대표적인 방법론. 라벨링이 되어 있지 않은 데이터들 내에서 **비슷한 특징이나 패턴을 가진 데이터들끼리 군집화**한 후, 새로운 데이터가 어떤 군집에 속하는지를 추론한다. K-means는 대표적인 Clustering의 알고리즘 중 하나.*

scikit-learn 패키지에서 제공.

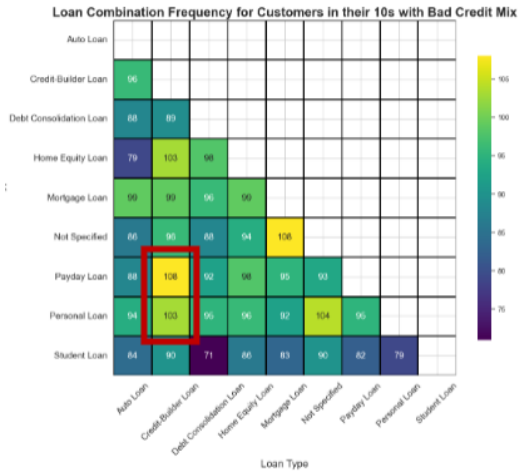
KDE plot이란?

- KDE(Kernel Density Estimation)plot은 데이터의 분포를 시각화 하는데 사용
- 주어진 데이터셋의 확률 밀도 함수를 추정하여 곡선형태로 그래프 출력
- X축이 -2부터 시작하는 이유는 KDE의 특성으로 주어진 데이터 포인트의 범위를 넘어선 영역까지 확장



히트맵(heatmap)이란?

데이터 값의 크기에 따라 색상을 다르게 표현하여 시각적으로 데이터를 쉽게 비교할 수 있게 해주는 그래프
그래프에서 각 셀의 색상은 특정 대출 조합의 빈도를 나타내고, 색상의 밝기나 채도에 따라 빈도의 크기를 직관적으로 파악



부록

대출 종류	설명
Auto Loan(자동차 대출)	**자동차 구매를 위한 대출**로, 자동차의 가격 일부 또는 전체를 대출로 받아 구매할 수 있습니다.
Student Loan(학자금 대출)	**학비와 관련된 경비를 지원**하기 위한 대출로, 학생들이 고등교육을 받는 데 사용됩니다.
Mortgage Loan(모기지 대출)	**주택을 구매하거나 주택 융자를 지원하기 위한** 대출로, 주택 구매 시 사용됩니다.
Credit-Builder Loan(신용 개선 대출)	**신용 점수를 높이기 위한 목적으**로 사용되는 대출로, 신용 기록을 개선하는 데 도움을 줍니다.
Not Specified(명시되지 않은 대출)	**구체적으로 지정되지 않은 대출 유형을 의미**합니다. 해당 대출의 세부 정보가 명확하지 않을 수 있습니다.

대출 종류	설명
Home Equity Loan(주택 자산 대출)	**주택 자산을** 담보로 한 대출로, 주택의 가치에 따라 대출을 받을 수 있습니다.
Payday Loan(연봉 대출)	**급여를 받는 날에 상환할 수 있도록 설계된 대출**로, 보통 긴급한 현금 필요 상황에서 사용됩니다.
Debt Consolidation(부채 통합 대출)	**다수의 채무를 하나의 대출로 합치는 대출**로, 관리 용이성을 위해 사용됩니다.
Personal Loan(개인 대출)	**개인적인 용도**로 사용할 수 있는 일반적인 대출로, 예를 들어 결혼, 여행 등에 사용될 수 있습니다.

Good

	Outstanding_Debt	Num_of_Delayed_Payment	Interest_Rate	Num_Bank_Accounts	Delay_from_due_date
count	3704.000000	3704.000000	3704.000000	3704.000000	3704.000000
mean	740.016998	7.112401	6.411177	2.882582	10.126035
std	431.603013	4.618913	3.438781	1.984402	6.903406
min	0.230000	0.000000	1.000000	0.000000	0.000000
25%	374.157500	3.000000	3.000000	1.000000	5.000000
50%	731.255000	7.000000	6.000000	3.000000	9.000000
75%	1119.190000	10.250000	9.000000	4.000000	14.000000
max	1498.020000	22.000000	12.000000	9.000000	31.666667

Standard

	Outstanding_Debt	Num_of_Delayed_Payment	Interest_Rate	Num_Bank_Accounts	Delay_from_due_date
count	5648.000000	5648.000000	5648.000000	5648.000000	5648.000000
mean	1066.768465	14.046388	15.224681	5.674575	19.347689
std	686.251420	3.820381	36.318880	1.826941	10.130078
min	0.340000	5.000000	1.000000	2.000000	0.000000
25%	499.565000	11.000000	9.000000	4.000000	11.666667
50%	1021.850000	14.000000	14.000000	6.000000	18.666667
75%	1447.250000	17.250000	19.000000	7.000000	25.666667
max	2699.170000	23.000000	2695.000000	11.000000	63.333333

Bad

	Outstanding_Debt	Num_of_Delayed_Payment	Interest_Rate	Num_Bank_Accounts	Delay_from_due_date
count	2920.000000	2920.000000	2920.000000	2920.000000	2920.000000
mean	3001.853592	19.976455	24.454110	7.964127	38.510103
std	1064.273867	3.228255	5.694514	1.407458	13.911897
min	1250.010000	12.500000	15.000000	5.000000	10.000000
25%	2138.542500	17.250000	19.000000	7.000000	26.750000
50%	2849.790000	20.000000	24.000000	8.000000	38.333333
75%	3885.677500	22.750000	29.000000	9.000000	50.666667
max	4998.070000	27.500000	34.000000	11.000000	67.000000

데이터 타입

```
Customer_ID                 object
Occupation                  object
Annual_Income              float64
Monthly_Inhand_Salary      float64
Credit_Mix                  object
Payment_of_Min_Amount       object
Payment_Behaviour           object
Num_Bank_Accounts          float64
Num_Credit_Card            float64
Interest_Rate              float64
Num_of_Loan                float64
Delay_from_due_date        float64
Num_of_Delayed_Payment     float64
Changed_Credit_Limit       float64
Num_Credit_Inquiries       float64
Outstanding_Debt           float64
Credit_Utilization_Ratio   float64
Total_EMI_per_month        float64
Amount_invested_monthly    float64
Monthly_Balance            float64
Age                        float64
Credit_History_Age           int64
Auto Loan                    int64
Credit-Builder Loan          int64
Debt Consolidation Loan      int64
Home Equity Loan             int64
Mortgage Loan                int64
None                         int64
Not Specified                int64
Payday Loan                  int64
Personal Loan                int64
Student Loan                 int64
dtype: object
```