아웃소싱 플랫폼
조사 보고서

IM Digital Banker Academy 5기
4조 이우태 정혜은 조예진 한민정

CONTENTS

배경 분석

[IT 아웃소싱 시장 급성장]

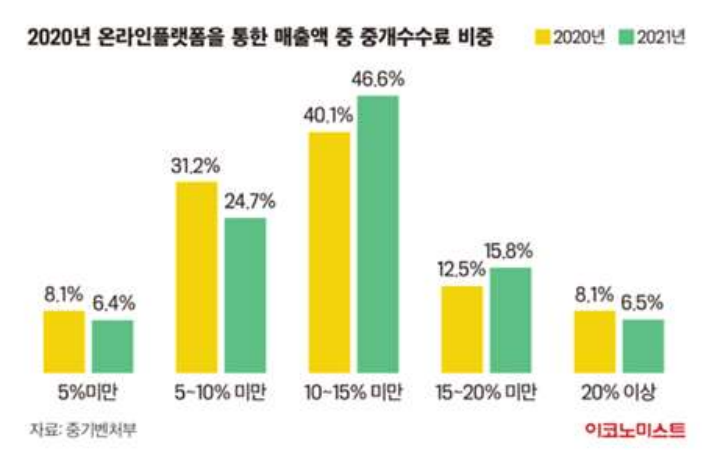
IT 기술 발전과 디지털 전환으로 기업과 스타트업에서

IT 전문가 수요 급증함.

아웃소싱 플랫폼을 통한 전문가 매칭 서비스의 중요성 증가

(예: 숨고, 누적 견적 40배 증가).

배경 분석



[L사의 수익구조]

L사는 수수료 기반 모델에 의존하고 있으며, 거래 성사율 감소 시 수수료 수익이 줄어드는 구조임. 따라서, 거래 성사율 개선 및 적정 수수료 책정 방안이 필요.

배경 분석

[수익 모델 다각화 필요]

타 기업을 참고한 결과 수수료 외에도 광고, 매칭 서비스, 부가 서비스 등 다양한 수익 모델 도입이 필요함.

기존 수익 구조의 한계를 극복하려면 비즈니스 채널 확장이 필요함.

[프로젝트 목표]

1년치 L 사의 거래 데이터를 분석해 거래 성사율에 영향을 미치는 주요 요인 분석, 적정 수수료에 대한 인사이트 도출.

또한, 고객 만족도, 전문가 특성, 가격 등을 고려해 최적화된 수익 모델을 도출하고, 수익 모델을 구축할 예정.

강점 (Strength)

L사의 아웃소싱 플랫폼은 2015년부터 시장에 자리 잡았으며, 2020년부터 분기별 안정적인 매출과 신규 가입자 증가로 **국내 최대 IT 아웃소싱 애플리케이션으로 성장했다**.

약점 (Weakness)

L사는 수수료 기반 수익 구조로 거래량 감소 시 매출 감소가 발생함.

거래 성사율 개선과 수수료 조정이 필요하며,

경쟁사들은 차별화된 서비스로 이를 해결하고 있다.

기회 (Opportunity)

아웃소싱 수요 급증에 따라 신규 고객 유치와 이탈 방지가 가능하며,

업계 경쟁력을 강화하면 더 많은 성장 기회를 만들 수 있다.

위협 (Threat)

경쟁사의 공격적인 마케팅으로 고객 이탈 및 신규 고객 확보 어려움.

서비스 외부 거래 전환으로 플랫폼 사용률 감소 우려.

불명확한 가격 체계로 가격 경쟁 심화,

L사는 체계적인 가격 구조가 필요.

프로젝트 주제

전문가의 특성과 서비스 유형이

거래 성사율 및 고객 만족도에 끼치는

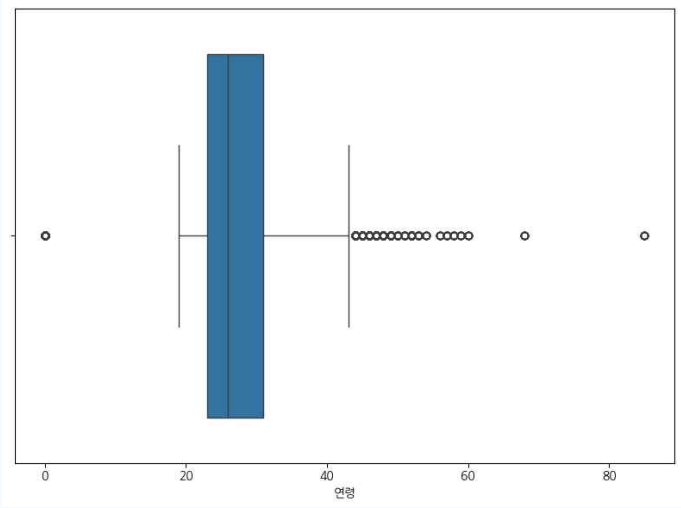
영향분석 및 가이드라인 제시

Customer_data 분포

	연령	서비스총구매수	총구매금액	총수정요청수	총추가결제금액
count	135722	137526	137521	137526	137526
mean	28	3	157699	0	10010
std	7	4	472974	2	18476
min	0	1	0	-6	0
25%	23	1	20200	0	0
50%	26	1	53700	0	4270
75%	31	3	132200	0	11500
max	85	96	45052500	120	1445200

고객 ID를 제외한 수치형의 기술통계

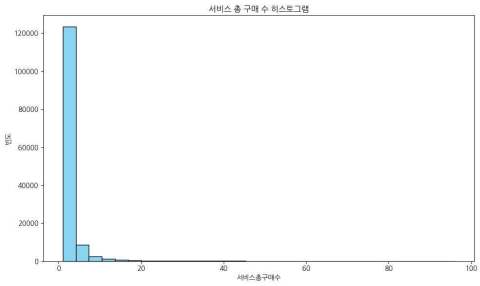
Customer_data 분포



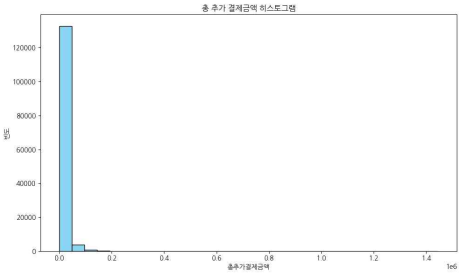
연령의 BoxPlot

연령 0, 85를 이상치라고 판단하고 제거함

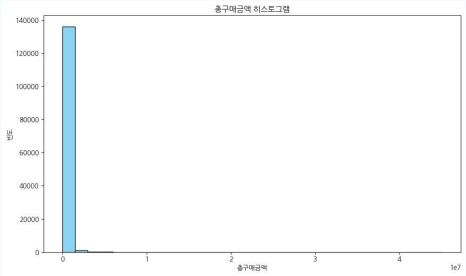
Customer_data 분포



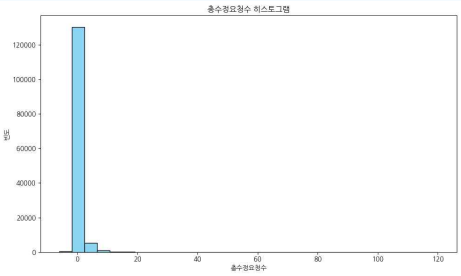
서비스 총 구매 수



총 추가 결제 금액



총 구매금액



총 수정 요청 수

Customer_data 분포

📌 기본 기술통계 정보

	count	unique	top	freq
사용기기	137526	243	iPhone	54958
사용OS	137526	45	Android8.0.0	49074
유입경로	135920	10	유튜브	52111
회원상태	137526	3	정상회원	132505
거주지	137526	7	경기도	48907
성별	127267	2	남성	93834

📌 추가 통계 정보

	고유값 개수	결측값 개수	최빈값	최빈값 개수
사용기기	243	0	iPhone	54958
사용OS	45	0	Android8.0.0	49074
유입경로	10	1606	유튜브	52111
회원상태	3	0	정상회원	132505
거주지	7	0	경기도	48907
성별	2	10259	남성	93834

[성별 결측치 제거]

결측치가 너무 많아 사전 단계에서 결측치을 미리 제거함.

[유입경로 결측치 처리]

데이터 수에 비해 결측치가 적어, isnull을 이용해 결측치를 제거함.

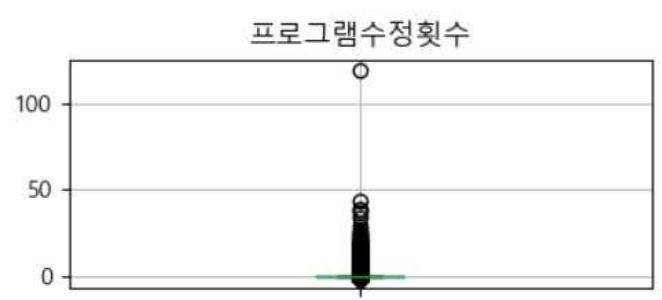
Log_Data02 분포

	프로그램수정횟수	추가결제금액	판매금액	서비스가격	평점	이용자수	총금액	매출
count	344299.00	344299.00	344278.00	344277.00	344277.00	344277.00	344278.00	344278.00
mean	0.17	3998.39	62992.30	60389.45	4.83	54.80	66990.61	465480.14
std	0.72	9074.91	256830.75	257547.78	0.81	90.53	257375.31	1799037.15
min	-1.00	0.00	0.00	5000.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	10000.00	9000.00	5.00	8.00	11000.00	79200.00
50%	0.00	0.00	23950.00	20000.00	5.00	22.00	30000.00	195000.00
75%	0.00	6500.00	50000.00	50000.00	5.00	62.00	53800.00	375700.00
max	119.00	1445200.00	44000000.00	44000000.00	5.00	584.00	44000000.00	286000000.00

거래일자, 거래 취소 일자, 고객 ID, 서비스명, 판매자, 서비스 번호를 제외한 수치형의 기술 통계

수수료율, 거래 취소 여부, 대분류는 범주형으로 따로 확인 필요함

Log_Data02 분포



프로그램 수정 횟수가 119인 인스턴스는 이상치라 판단하고 제거함

	수수료율	거래취소여부	대분류
count	344299.00	344299.00	344277
unique	3.00	2.00	12
top	6.50	0.00	홈페이지
freq	241009.00	339702.00	124007

거래 취소 여부의 결측치를 0으로 채움

Service_Data 분포

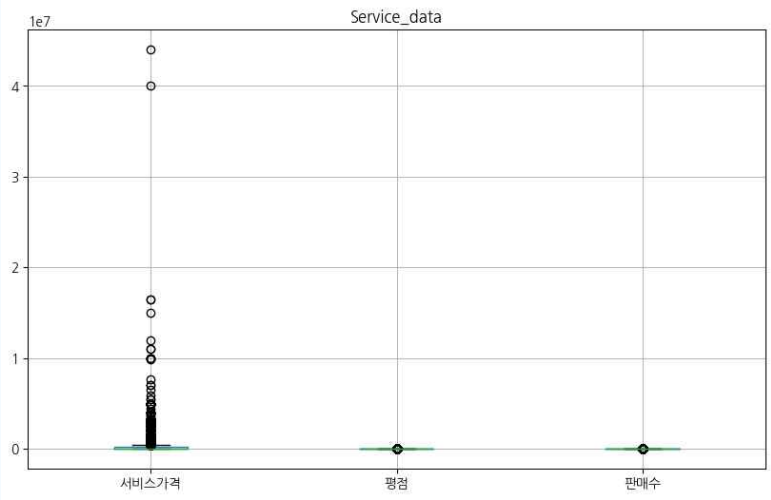
	서비스가격	평점	판매수
count	4158.00	4158.00	4159.00
mean	283340.79	4.11	82.78
std	1251369.55	1.88	199.83
min	5000.00	0.00	1.00
25%	15000.00	4.90	5.00
50%	50000.00	5.00	16.00
75%	200000.00	5.00	68.00
max	44000000.00	5.00	3007.00

수치형의 기술통계

	서비스명	판매자	대분류
count	4158	4158	4158
unique	2787	2066	12
top	반응형 웹사이트 제작해 드립니다.	TNIX	홈페이지
freq	12	16	1485

범주형 데이터의 기술통계

Service_Data 분포



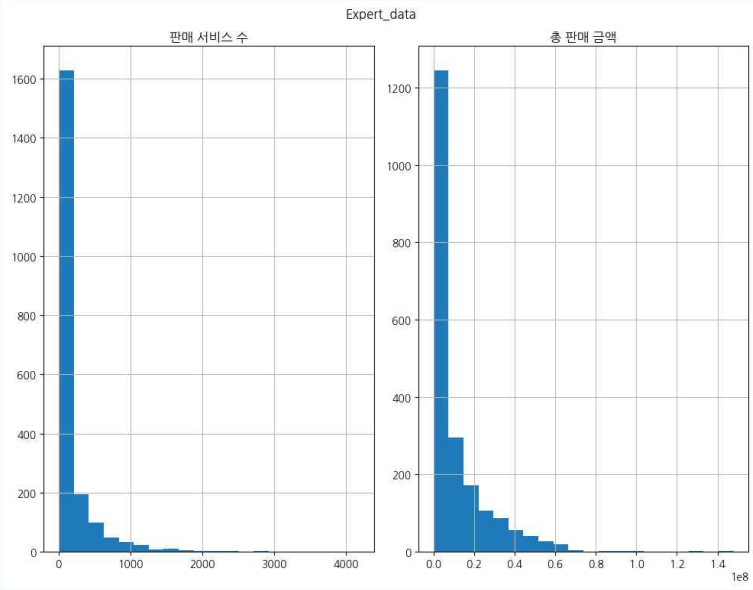
서비스 가격, 평점, 판매수의 데이터 분포

서비스 가격은 단순 가격의 분포라 이상치라 판단하기 어려워 따로 제거하지 않음

Expert_Data 분포

	판매 서비스 수	총 판매 금액
count	2066.00	2066.00
mean	166.64	11163308.06
std	342.31	16184984.28
min	1.00	5000.00
25%	9.00	961850.00
50%	36.00	4139370.00
75%	158.75	15084200.00
max	4180.00	147927420.00

판매자 ID를 제외한 수치형의 기술통계



판매 서비스 수, 총 판매 금액의 분포

Expert_Data 분포

	판매자ID	판매자	프리미엄 서비스 가입여부	신속 알람 서비스 사용여부
count	2066	2066	2066	2066
unique	2066	2066	2	2
top	V2_0	0to1	미가입	미사용
freq	1	1	1774	1687

Category형의 데이터 기술통계

전체 가설 소개

1. 전문가의 특성이 고객 만족도에 영향을 미칠 것이다.

2. 가격, 수수료에 따라 거래 성사율에 차이가 있을 것이다.

3. 수수료 비율과 고객 유입(신규 가입자 수) 간에 관계가 있을 것이다.

4. 고객 만족도, 전문가 평가 점수, 수수료 등이 거래 성사 여부에 영향을 미칠 것이다.

5. 시간대별(주/야) 거래 성사율에 차이가 있을 것이다.

6. 고객의 이탈율과 신규 가입 고객 간에 관계가 있을 것이다.

데이터 전처리

- 각각 필요한 요소를 하나씩 가져다 쓰는 게 아닌 한 테이블 안에서 처리할 수 있도록 전처리함 -

•데이터 병합 목표

총 4개의 데이터를 통합하여 하나의 데이터 테이블로 만듦

각 테이블의 결측치와 이상치를 모두 수정하여 데이터의 무결성과 정확성 유지

•병합 방법

총 3번의 merge 실행

병합 기준: 고객 ID, 판매자 이름, 서비스 번호

•데이터 처리

모든 결측치 및 이상치 수정

수정 사항이 다른 테이블에 영향을 주지 않도록 처리

데이터 이슈 파악

1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)

1-1. 프로그램 수정 횟수 이슈 (음수 및 소수)

프로그램 수정 횟수가 음수(-1) 및 소수(0.3) 값이 포함됨

음수: 7737개 / 344366개 데이터 (약 2%)

소수: 18개 / 344266개 데이터 (약 0.005%) 로 비율이 작아 값을 제거함

수정 횟수 -1인 데이터의 서비스 판매 금액이 0으로 확인되어,

이 데이터는 거래 취소 여부를 1로 설정 (1->취소됨, 0->정상거래)

최종적으로, 프로그램 수정 횟수를 범주형으로 설정

데이터 이슈 파악

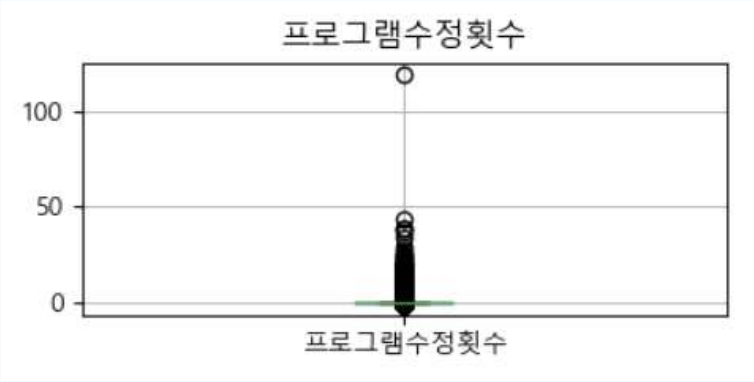
1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)

거래일자	고객ID	프로그램수정횟수	거래취소여부	서비스명	판매자	판매금액
20210201	8.01E+10	-1		주말작업가능	궁봇	0
20210201	4.01E+10	-1		웹사이트 기	greydoll	0
20210201	1E+10	-1	1	고퀄리티 반	포인원	0
20210201	1.3E+11	-1	1	사람 일을 다	오토프로머	0
20210201	1.3E+10	-1	1	사람 일을 다	오토프로머	0
20210201	1.02E+10	-1	1	1일이내 작업	var	0

거래일자	고객ID	프로그램수정횟	거래취소여부	서비스명	판매자	판매금액
20210201	8.01E+10	-1	1	주말작업가	궁봇	0
20210201	4.01E+10	-1	1	웹사이트 기	greydoll	0
20210201	1E+10	-1	1	고퀄리티 반	포인원	0
20210201	1.3E+11	-1	1	사람 일을	오토프로머	0
20210201	1.3E+10	-1	1	사람 일을	오토프로머	0
20210201	1.02E+10	-1	1	1일이내 작	var	0

데이터 이슈 파악

1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)



1-2. 프로그램 수정 횟수 이슈 (이상치)

프로그램 수정 횟수 119인 데이터가 1개 존재

다른 값들의 분포는 0 ~ 44로, 119는 이상치로 판단되어 제거함

데이터 이슈 파악

1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)

1-3. 데이터 컬럼 이슈 (총 금액 문제)

```
import pandas as pd
df1 = pd.read_csv('Log_Data02.csv')
df4 = pd.read_csv('Expert_Data.csv')
```

```
df1['총판매금액'] = df1['서비스가격'] + df1['추가결제금액'] + df1['추가결제금액']
```

```
df4 = df4[['판매자', '총 판매 금액']].drop_duplicates()
merged = pd.merge(df1, df4, on='판매자', how='inner')
merged['총판매금액_일치여부'] = (df1_merged['총판매금액'] == merged['총 판매 금액'])
result = merged[['판매자', '총판매금액', '총 판매 금액', '총판매금액_일치여부']].drop_duplicates()
result['총판매금액_일치여부'].value_counts()
```

```
총판매금액_일치여부
True    63920
Name: count, dtype: int64
```

"총 판매 금액 = Σ(서비스 가격 + 추가 결제 금액 + 추가 결제 금액)" 으로 되어 있는 것을 확인

데이터 이슈 파악

1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)

1-3. 데이터 컬럼 이슈 (총 금액 문제)

```
import pandas as pd
df_Log = pd.read_csv('Log_Data02.csv')
df_Log['총금액'] = df_Log['판매금액'] + df_Log['추가결제금액']
df_Log['매출'] = df_Log['총금액'] * df_Log['수수료율']
df_Log.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344299 entries, 0 to 344298
Data columns (total 17 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   거래일자       344299 non-null  int64
 1   수수료율       344299 non-null  float64
 2   고객ID       344299 non-null  int64
 3   프로그램수정횟수   344299 non-null  float64
 4   추가결제금액     344299 non-null  int64
 5   거래취소여부      4597 non-null    float64
 6   거래취소일자      4597 non-null    float64
 7   서비스명        344277 non-null  object
 8   판매자         344277 non-null  object
 9   판매금액        344278 non-null  float64
 10  서비스가격       344277 non-null  float64
 11  평점          344277 non-null  float64
 12  이용자수        344277 non-null  float64
 13  대분류         344277 non-null  object
 14  서비스번호       344277 non-null  float64
 15  총금액         344278 non-null  float64
 16  매출          344278 non-null  float64
dtypes: float64(11), int64(3), object(3)
memory usage: 44.7+ MB
```

- Log 데이터 가격 이슈 처리 -

"총 금액 = Σ (판매금액 + 추가 결제 금액)",

"판매금액 = 서비스 가격 + 추가 결제 금액" 으로

정의하고 Log 데이터에 컬럼 추가

데이터 이슈 파악

1. Log_Data_2 이슈 (merge 하기 전 최종 Log_Data 데이터)

1-4. 평점 이슈

평점이 각 고객이 서비스에 대해 남긴 평점이 아닌 기존 서비스에 대한 평점이 기입되어 있는 것을 확인

고객ID	프로그	추가결	거래취	거래취	서비스명	판매자	판매금	서비스	평점	이용자	대분류	서비스
101010422	0	11200			퍼블리싱 HTML, CS	네오웨이	41200	30000	5	6	홈페이지	3614
101010422	0	11200			WEB 반응형 퍼블리	db퍼블	26200	15000	5	6	홈페이지	3611
101010422	0	11200			각종 정적, 동적 사이	마슬링	21200	10000	5	28	데이터	2670
101010422	0	11200			워드프레스 버그 수	워드넷	21200	10000	4.9	69	홈페이지	2343

따라서, 고객과 평점을 같이 보는 것이 아닌 **평점을 별개의 요소로 판단해야 하는 것을 확인**

데이터 이슈 파악

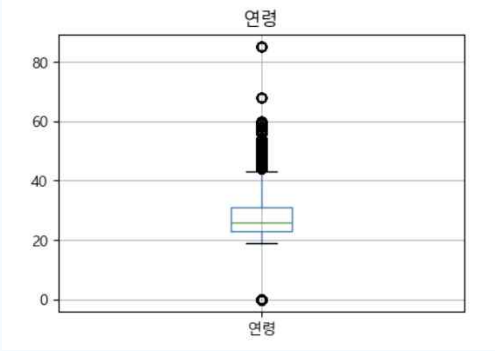
2. Customer_Data 이슈

2-1. 연령 이슈

연령이 0인 데이터가 존재 (약 0.1%) → 논리적으로 불가능

BoxPlot을 통해 확인한 결과, 연령 0은 매우 떨어진 값으로 이상치로 판단되어 제거

연령 85인 데이터 1개 존재 → IQR을 기준으로 매우 떨어진 값으로 이상치로 판단되어 제거



데이터 이슈 파악

3. Service_Data 이슈

3-1. 판매수 이슈

동일한 서비스명에 여러 개의 판매수가 중복 분포 되어 있는 것을 확인

중복 값을 다 더하여 1개의 값으로 정리함

A	B	C	D	E	F	G
서비스명	판매자	서비스가격	평점	대분류	서비스번호	판매수
17년경력 VBA 이용하여 매크로 레포트	EBConsulting	30000	5	기타	2233	9
17년경력 VBA 이용하여 매크로 레포트	EBConsulting	30000	5	기타	2233	82
17년경력 VBA 이용하여 매크로 레포트	EBConsulting	30000	5	기타	2233	711
17년경력 VBA 이용하여 매크로 레포트	EBConsulting	30000	5	기타	2233	23

서비스명	판매자	서비스가격	평점	대분류	서비스번호	판매수
17년경력 VBA 0	EBConsulting	30000	5	기타	2233	825

데이터 이슈 파악

3. Service_Data 이슈

3-2. 평점 소수점 이슈

Service 데이터와 Log 데이터를 merge한 결과 평점이 같은 4.9이나 상세 데이터를 통해 비교한 결과

평균 평점이 4.9와 4.8999로 되어 있는 것을 확인

```
import pandas as pd
df_service=pd.read_csv('Service_Data.csv')
df_log=pd.read_csv('Log_Data02.csv')
df_service.info()
df_log.info()
```

```
grouped_stats1 = df_service.groupby(['서비스명', '판매자', '서비스번호'])['평점'].describe()
grouped_stats2 = df_log.groupby(['서비스명', '판매자', '서비스번호'])['평점'].describe()
grouped_stats1 = grouped_stats1.drop(['count','std'],axis=1)
grouped_stats2 = grouped_stats2.drop(['count','std'],axis=1)
```

```
# 1. 값이 정확히 같은지 전체 비교 (DataFrame 단위)
전체_동일 = grouped_stats1.equals(grouped_stats2)
print("전체 값이 완전히 같은가? →", 전체_동일)
```

```
전체 값이 완전히 같은가? → False
```

```
[80]: df_service_df['mean'][4]

[80]: 4.9

[81]: df_log_df['mean'][4]

[81]: 4.899999999999998
```

데이터 이슈 파악

3. Service_Data 이슈

3-2. 평점 소수점 이슈

Round 처리하여 소수점을 반올림해 **평균 평점을 같은 4.9로 인식할 수 있도록 설정**

```
차이 나는 행 수: 0
Empty DataFrame
Columns: [mean_service, min_service, 25%_service, 50%_service, 75%_service, max_service, mean_log, min_log, 25%_log, 50%_log, 75%_log, max_log]
Index: []
```

```
grouped_stats1.round(1).equals(grouped_stats2.round(1))
```

```
True
```

가설 검증 불가능 원인

[1. 전문가의 특성이 고객 만족도에 영향을 미칠 것이다.] 의 불가능 원인

grouped_stats1

서비스명	판매자	서비스번호	mean	min	25%	50%	75%	max
"P2P 대출" 서비스를 위한 기획을 해 드립니다.	디자인쇼크리더	5651.0	0.0	0.0	0.0	0.0	0.0	0.0
"개발 공부 전 필독" IT 서비스 기본 개념 Ebook을 드립니다.	DDAAC	3361.0	4.8	4.8	4.8	4.8	4.8	4.8
"디자인 에이전시" UXUI 화면 설계, 기획안 제작해 드립니다.	VOXDesign	4775.0	5.0	5.0	5.0	5.0	5.0	5.0
"디자인적인" 퍼블리싱 제작해 드립니다.	하디HADE	2984.0	5.0	5.0	5.0	5.0	5.0	5.0
"반응형 웹 퍼블리싱" 모든 퍼블리싱 작업 맞춤 진행	ThunDesign	6122.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
회로설계 및 PCB 설계 및 제작 해 드립니다.	해줌랩	2826.0	5.0	5.0	5.0	5.0	5.0	5.0
회사 도메인으로 그룹웨어 메일 수신 설정 도와 드립니다.	그린티라떼좋아	3871.0	5.0	5.0	5.0	5.0	5.0	5.0
회사업무관리 ERP, MES,WMS 프로그램 개발	항상초심	4999.0	5.0	5.0	5.0	5.0	5.0	5.0
효과적인 연락처 수집용 랜딩페이지 제작해 드립니다.	안단테연구소	4911.0	5.0	5.0	5.0	5.0	5.0	5.0
희망하는 웹사이트의 정보를 수집웹크롤링해 드립니다.	제임슈	4160.0	5.0	5.0	5.0	5.0	5.0	5.0

2863 rows × 6 columns

서비스 데이터 평점 기준으로 기술 통계 출력

grouped_stats2

서비스명	판매자	서비스번호	mean	min	25%	50%	75%	max
"P2P 대출" 서비스를 위한 기획을 해 드립니다.	디자인쇼크리더	5651.0	0.0	0.0	0.0	0.0	0.0	0.0
"개발 공부 전 필독" IT 서비스 기본 개념 Ebook을 드립니다.	DDAAC	3361.0	4.8	4.8	4.8	4.8	4.8	4.8
"디자인 에이전시" UXUI 화면 설계, 기획안 제작해 드립니다.	VOXDesign	4775.0	5.0	5.0	5.0	5.0	5.0	5.0
"디자인적인" 퍼블리싱 제작해 드립니다.	하디HADE	2984.0	5.0	5.0	5.0	5.0	5.0	5.0
"반응형 웹 퍼블리싱" 모든 퍼블리싱 작업 맞춤 진행	ThunDesign	6122.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
회로설계 및 PCB 설계 및 제작 해 드립니다.	해줌랩	2826.0	5.0	5.0	5.0	5.0	5.0	5.0
회사 도메인으로 그룹웨어 메일 수신 설정 도와 드립니다.	그린티라떼좋아	3871.0	5.0	5.0	5.0	5.0	5.0	5.0
회사업무관리 ERP, MES,WMS 프로그램 개발	항상초심	4999.0	5.0	5.0	5.0	5.0	5.0	5.0
효과적인 연락처 수집용 랜딩페이지 제작해 드립니다.	안단테연구소	4911.0	5.0	5.0	5.0	5.0	5.0	5.0
희망하는 웹사이트의 정보를 수집웹크롤링해 드립니다.	제임슈	4160.0	5.0	5.0	5.0	5.0	5.0	5.0

2863 rows × 6 columns

로그 데이터 평점 기준으로 기술 통계 출력

가설 검증 불가능 원인

[1. 전문가의 특성이 고객 만족도에 영향을 미칠 것이다.] 의 불가능 원인

[13]:
```
# 2. 값이 다른 부분 찾기
차이나는_행 = (grouped_stats1.round(1) != grouped_stats2.round(1))

# 3. 하나라도 다른 행만 추출
diff_indices = 차이나는_행.any(axis=1)
차이나는_데이터 = pd.concat([
    grouped_stats1[diff_indices].add_suffix('_service'),
    grouped_stats2[diff_indices].add_suffix('_log')
], axis=1)

print("차이 나는 행 수:", diff_indices.sum())
print(차이나는_데이터)
```

```
차이 나는 행 수: 0
Empty DataFrame
Columns: [mean_service, min_service, 25%_service, 50%_service, 75%_service, max_service, mean_log, min_log, 25%_log, 50%_log, 75%_log, max_log]
Index: []
```

[14]:
```
grouped_stats1.round(1).equals(grouped_stats2.round(1))
```

[14]:
```
True
```

서비스 데이터의 평점과 로그 데이터의 평점을 서로 비교했더니 **모두 다 같은 것을 확인**

가설 검증 불가능 원인

[2. 가격, 수수료에 따라 거래 성사율에 차이가 있을 것이다.] 의 불가능 원인

	대분류	거래성사율
0	UnReal	97.410072
1	게임	98.954821
2	기타	98.320618
3	기획	97.839667
4	데이터	97.812754
5	디자인	95.684667
6	모바일	97.870020
7	인공지능	97.105644
8	커머스	98.177912
9	프로그래밍	97.781255
10	프로그램	96.832449
11	홈페이지	97.650704

서비스 대분류(12개) 별로 거래 성사율(mean)을 계산

거래 성사율 = (거래 성사 수/전체 거래 수) * 100

12건 확인 결과, 최저 약 95% ~ 최고 약 98%로

거래 성사율 지표의 분포가 고르지 않음을 확인

가설 검증 불가능 원인

[3. 수수료 비율과 고객 유입(신규 가입자 수) 간에 관계가 있을 것이다.] 의 불가능 원인

```
[10]: merge_log_cus.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 335782 entries, 0 to 335781
Data columns (total 24 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   거래일자       335782 non-null  int64
 1   수수료율       335782 non-null  float64
 2   고객ID       335782 non-null  int64
 3   프로그램수정횟수   335782 non-null  float64
 4   추가결제금액     335782 non-null  int64
 5   거래취소여부     335782 non-null  float64
 6   서비스명       335782 non-null  object
 7   판매자        335782 non-null  object
 8   판매금액       335782 non-null  float64
 9   서비스가격      335782 non-null  float64
 10  평점         335782 non-null  float64
 11  이용자수       335782 non-null  float64
 12  대분류        335782 non-null  object
 13  서비스번호      335782 non-null  float64
 14  사용기기       335782 non-null  object
 15  사용OS       335782 non-null  object
 16  유입경로       335782 non-null  object
 17  회원상태       335782 non-null  object
 18  거주지        335782 non-null  object
 19  연령         335782 non-null  float64
 20  서비스총구매수    335782 non-null  int64
 21  총구매금액      335782 non-null  float64
 22  총수정요청수     335782 non-null  float64
 23  총추가결제금액    335782 non-null  int64
dtypes: float64(11), int64(5), object(8)
```

수수료 비율과 신규 고객 간에 관계를 확인하기 위하여

‘Log_Data’ 와 ‘Customer_Data’ 데이터를 결합해

데이터 분포를 확인

수수료 비율은 있지만 **신규 고객에 대한 컬럼 정보가 없음**을 확인

가설 검증 불가능 원인

[4. 고객 만족도, 전문가 평가 점수, 수수료 등이 거래 성사 여부에 영향을 미칠 것이다.] 의 불가능 원인

	대분류	거래성사율
0	UnReal	97.410072
1	게임	98.954821
2	기타	98.320618
3	기획	97.839667
4	데이터	97.812754
5	디자인	95.684667
6	모바일	97.870020
7	인공지능	97.105644
8	커머스	98.177912
9	프로그래밍	97.781255
10	프로그램	96.832449
11	홈페이지	97.650704

거래 성사율 지표의 **분포가 편향**되어 있어

고객 만족도, 전문가 평가 점수,

수수료 등의 영향과 의미 없음을 확인

가설 검증 불가능 원인

[5. 시간대별(주/야) 거래 성사율에 차이가 있을 것이다.] 의 불가능 원인

```
: df_log.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344265 entries, 0 to 344264
Data columns (total 14 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   거래일자       344265 non-null  int64
 1   수수료율       344265 non-null  float64
 2   고객ID       344265 non-null  int64
 3   프로그램수정횟수   344265 non-null  float64
 4   추가결제금액     344265 non-null  int64
 5   거래취소여부     344265 non-null  float64
 6   서비스명       344265 non-null  object
 7   판매자        344265 non-null  object
 8   판매금액       344265 non-null  float64
 9   서비스가격      344265 non-null  float64
 10  평점         344265 non-null  float64
 11  이용자수       344265 non-null  float64
 12  대분류        344265 non-null  object
 13  서비스번호      344265 non-null  float64
dtypes: float64(8), int64(3), object(3)
memory usage: 36.8+ MB
```

거래 취소 여부를 거래 취소 -> 1, 정상 거래 -> 0 로 나타내어

$$\text{거래성사율} = \left(\frac{\text{정상거래수}}{\text{전체거래수}}\right) * 100$$ 으로 나타내고자 함

시간대에 대한 데이터 정보 존재하지 않은 것을 확인

가설 검증 불가능 원인

[6. 고객의 이탈율과 신규 가입 고객 간에 관계가 있을 것이다.] 의 불가능 원인

```
df_cus.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 134136 entries, 0 to 134135
Data columns (total 11 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   고객ID      134136 non-null  int64
 1   사용기기      134136 non-null  object
 2   사용OS      134136 non-null  object
 3   유입경로      134136 non-null  object
 4   회원상태      134136 non-null  object
 5   거주지       134136 non-null  object
 6   연령        134136 non-null  float64
 7   서비스총구매수  134136 non-null  int64
 8   총구매금액     134136 non-null  float64
 9   총수정요청수    134136 non-null  float64
 10  총추가결제금액  134136 non-null  int64
dtypes: float64(3), int64(3), object(5)
memory usage: 11.3+ MB
```

고객 이탈율 = $\frac{\text{len}(\text{고객상태}(\text{탈퇴} + \text{탈퇴진행중}))}{\text{len}(\text{고객ID})} \times 100$ 으로

나타내고자 함

신규 가입 고객에 대한 데이터가 존재하지 않은 것을 확인

최종 가설

- 거래 성사율이 편향되어 있기 때문에 거래 성사율이 아닌 거래 취소율로 대체하여 분석 -

거래 취소율 = $\frac{1(\text{취소여부})}{\text{전체 } Log \text{ 데이터 수}}$ x 100

판매금액 = 서비스 가격 + 추가 결제금액

총 금액= Σ (판매금액 + 추가 결제금액)

매출 = Σ(총금액 X 수수료율)

판매 서비스 수 = 판매자 별 판매 수를 더함

최종 가설

1. 서비스 가격과 추가 결제 금액 발생 여부가 서비스 번호 별 거래 취소율에 영향이 있을 것이다.

종속 변수 : 거래 취소율

독립 변수 : 서비스 가격, 추가 결제 금액 발생 여부

2. 수수료율(6.5/7.5/9)에 따라 / 거래 취소율에 차이가 있을 것이다.

종속 변수 : 거래 취소율

독립 변수 : 수수료율(6.5/7.5/9)

최종 가설

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무

종속 변수 : 거래 취소율

독립 변수 : 범주화 된 서비스 가격

4. 서비스의 특성이 판매 서비스 수에 미치는 영향

종속 변수: 판매 서비스 수

독립 변수: 프리미엄 서비스 가입 여부, 신속 알람 서비스, 서비스 평점, 서비스 가격, 대분류

1. 서비스 가격과 추가 결제 금액 발생 여부가 서비스 번호 별 거래 취소율에 영향이 있을 것이다.(t 검정)

귀무가설(H0) : 서비스 가격과 추가 결제 금액의 발생여부가 서비스 번호 별 거래취소율에 미치는 영향이 없다

대립가설(H1) : 서비스 가격과 추가 결제 금액의 발생여부가 서비스 번호 별 거래취소율에 미치는 영향이 있다

```
from scipy.stats import pearsonr

variables = ['서비스가격', '추가결제금액발생여부']
for var in variables:
    corr, p = pearsonr(df_5[var], df_5['거래취소율'])
    print(f"{var} ↔ 거래취소율 | 상관계수: {corr:.3f} | p-value: {p:.5f}")
```

```
서비스가격 ↔ 거래취소율 | 상관계수: 0.001 | p-value: 0.45541
추가결제금액발생여부 ↔ 거래취소율 | 상관계수: -0.012 | p-value: 0.00000
```

서비스 가격은 p-value가 0.455 > 0.05로 **거래 취소율과 영향이 없음**

추가 결제 금액 여부는 p-value가 0.00 < 0.05로 **거래 취소율과 영향이 있음**

1. 서비스 가격과 추가 결제 금액 발생 여부가 서비스 번호 별 거래 취소율에 영향이 있을 것이다.(t 검정)

추가 결제 금액 발생 여부(0 또는 1)에 따라 **거래 취소율의 평균 차이가 유의미한지 검정**하기 위해
독립 표본 t-검정 진행함

```
t-통계량: 6.7684, p-값: 0.0000
추가결제금액발생여부에 따른 거래취소율 차이가 통계적으로 유의미함.
```

추가결제금액이 발생한 경우(1)와 발생하지 않은 경우(0)의
거래취소율 평균이 통계적으로 유의미하게 차이가 남

2. 수수료율(6.5/7.5/9)에 따라 / 거래 취소율에 차이가 있을 것이다.

	대분류	거래취소율	판매금액	서비스가격	추가결제금액	총금액	매출
0	UnReal	2.589928	56500	55000	0	58000	412500
1	게임	1.045179	11150	5000	0	13800	97500
2	기타	1.679382	20000	11000	1100	25000	162500
3	기획	2.160333	49000	49000	1200	49000	318500
4	데이터	2.187246	20000	20000	0	25000	175500
5	디자인	4.315333	33000	33000	1200	36000	259600
6	모바일	2.129980	51400	50000	0	53800	391500
7	인공지능	2.894356	99000	99000	0	99000	643500
8	커머스	1.822088	10000	5000	0	13000	90000
9	프로그래밍	2.218745	28100	20000	0	30000	215800
10	프로그램	3.167551	50000	50000	0	50000	325000
11	홈페이지	2.349296	25000	20000	0	30000	195000

[서비스 대분류별 거래취소율]

$$\text{거래 취소율} = \frac{1(\text{취소여부})}{\text{전체 } Log \text{ 데이터 수}} \times 100$$

서비스별 거래취소율: 0% ~ 100%까지 다양한 분포

서비스 대분류별 거래취소율: 약 1% ~ 5%로 고르게 분포

거래 성사율: 약 95% ~ 최고 98%로 분포가 고르지 않음

거래취소율에 영향을 미치는 요인을 분석할 필요성 존재

따라서, 거래성사율을 거래취소율로 변경하여 분석 진행

2. 수수료율(6.5/7.5/9)에 따라 / 거래 취소율에 차이가 있을 것이다. (ANOVA 분산 분석)

귀무가설(H0): 수수료율에 따라 거래 취소율에 차이가 없다.

대립가설(H1): 수수료율에 따라 거래 취소율에 차이가 있다.

```python
# 수수료율 - 거래 취소율 간 아노바분석
group_6_5 = df[df['수수료율_범주'] == '6.5%']['거래취소율']
group_7_5 = df[df['수수료율_범주'] == '7.5%']['거래취소율']
group_9 = df[df['수수료율_범주'] == '9%']['거래취소율']

# ANOVA 분석: 세 그룹 간 거래 취소율 차이가 있는지 확인
f_statistic, p_value = stats.f_oneway(group_6_5, group_7_5, group_9)

print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")

# 유의미한 차이가 있는지 판단
alpha = 0.05  # 95% 신뢰구간
if p_value < alpha:
    print("수수료율 범주에 따라 거래취소율에 유의미한 차이가 있습니다.")
else:
    print("수수료율 범주에 따라 거래취소율에 유의미한 차이가 없습니다.")
```

```
F-statistic: 1.7673654608478784
P-value: 0.17078391808095975
수수료율 범주에 따라 거래취소율에 유의미한 차이가 없습니다.
```

도출된 P-value가 0.17 > 0.05 로 수수료율과 거래취소율 간에 유의미한 차이가 없다는 것을 확인함.

2-1. 수수료율(6.5/7.5/9)에 따라 / 매출에 관련이 있을 것이다. (ANOVA 분산 분석)

귀무가설(H0): 수수료율에 따라 매출에 차이가 없다.

대립가설(H1): 수수료율에 따라 매출에 차이가 있다.

```python
# 수수료율 - 매출 간 아노바분석 => 유의미한 차이 있음 확인

group_6_5 = df[df['수수료율_범주'] == '6.5%']['매출']
group_7_5 = df[df['수수료율_범주'] == '7.5%']['매출']
group_9 = df[df['수수료율_범주'] == '9%']['매출']

# ANOVA 분석: 세 그룹 간 매출 차이가 있는지 확인
f_statistic, p_value = stats.f_oneway(group_6_5, group_7_5, group_9)

print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")

# 유의미한 차이가 있는지 판단
alpha = 0.05  # 95% 신뢰구간
if p_value < alpha:
    print("수수료율 범주에 따라 매출에 유의미한 차이가 있습니다.")
else:
    print("수수료율 범주에 따라 매출에 유의미한 차이가 없습니다.")
```

```
F-statistic: 141.47279174684977
P-value: 3.846089337885e-62
수수료율 범주에 따라 매출에 유의미한 차이가 있습니다.
```

```
수수료율과 매출 간 상관계수: 0.02897849354112081
p-value: 2.630768202335032e-63
```

F-statistic 보면 **수수료율 집단 간 매출 차이가 상당이 크다는 것 알 수 있음.**

p-value 보면 0에 매우 가까운 값으로

수수료율과 매출 간 연관이 있을 것이란 해당 가설은 통계적으로 매우 유의미함

2-1. 수수료율(6.5/7.5/9)에 따라 / 매출에 관련이 있을 것이다.

Tukey's HSD 테스트로 사후분석 진행

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
6.5%	7.5%	60517.8454	0.0	41980.9089	79054.7819	True
6.5%	9%	169143.8975	0.0	144438.9622	193848.8328	True
7.5%	9%	108626.0521	0.0	80320.6314	136931.4728	True

수수료율이 높을수록 매출이 증가하는 경향이 있음

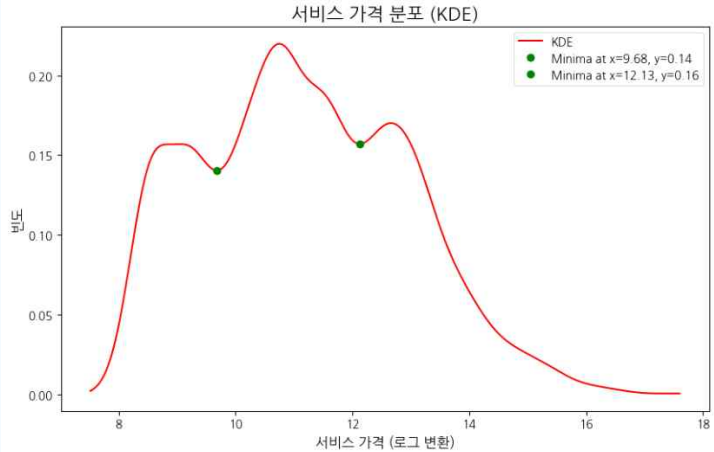
F-statistic 분석 결과, 세 범주 간에 매출 차이가 뚜렷하게 존재

특히, 6.5%와 9% 사이의 평균 매출 차이가 가장 큼

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무

원본 데이터 : Service_Data_0 (판매수만 전처리한 서비스 데이터)

전처리 데이터 : Merge_2 (log, service, expert, customer 4개를 모두 병합한 데이터)



원본 데이터 서비스 가격 분포 KDE 그래프

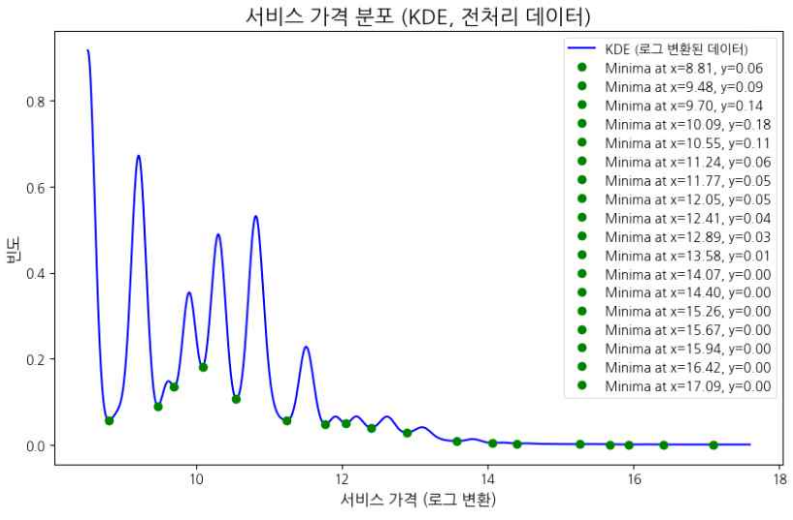
선형 회귀 모델링 및 통계적 검증

iM뱅크

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무

원본 데이터 : Service_Data_0 (판매수만 전처리한 서비스 데이터)

전처리 데이터 : Merge_2 (log, service, expert, customer 4개를 모두 병합한 데이터)



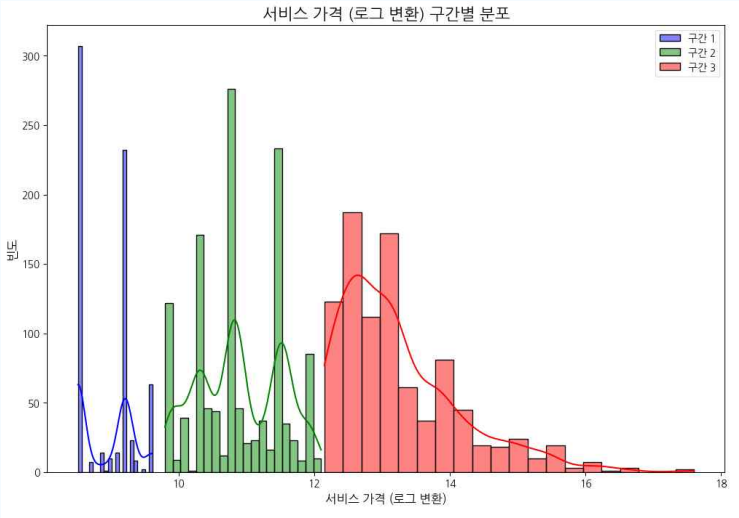
전처리 데이터 서비스 가격 분포 KDE 그래프

변곡점이 많아 기준점을 잡기 어려움

따라서, 원본 데이터를 통해

서비스 가격의 범주를 나누기로 결정함

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무

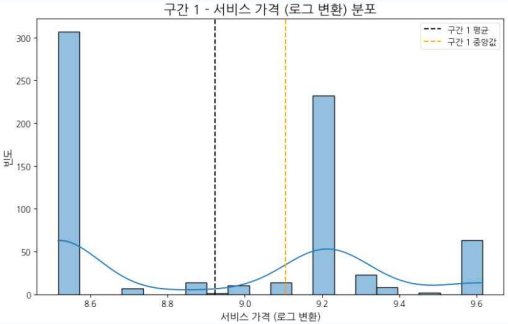


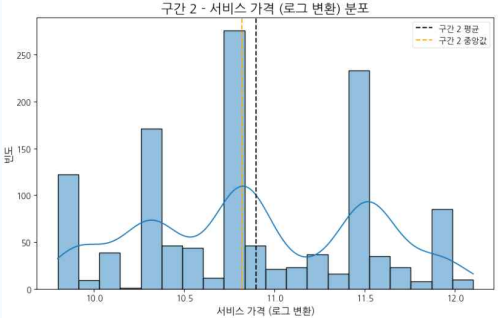
기준점1 = 15946.74 (서비스 가격)

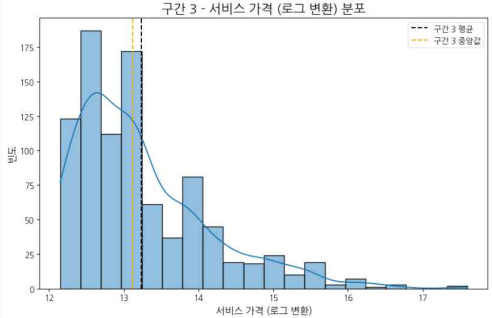
기준점2 = 185259.20 (서비스 가격)

원본 데이터의 3개의 기준점을 바탕으로 전처리 데이터의 가격 분포를 나눔

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무







구간 별로 각 서비스 가격의 분포

구간 1 : 5000 ~ 15946.74

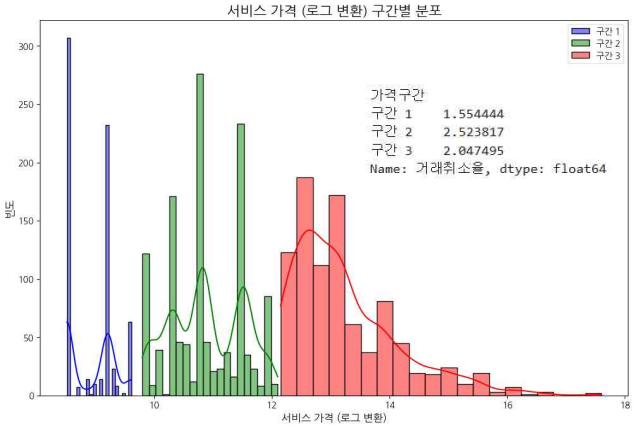
구간 2 : 15946.75 ~ 185259.19

구간 3 : 185259.20 ~ 44000000

선형 회귀 모델링 및 통계적 검증

iM뱅크

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무



ANOVA 분석 결과:

	sum_sq	df	F	PR(>F)
가격구간	426.579947	2.0	5.128195	0.005982
Residual	118910.451482	2859.0	NaN	NaN

Tukey HSD 사후 분석 결과:

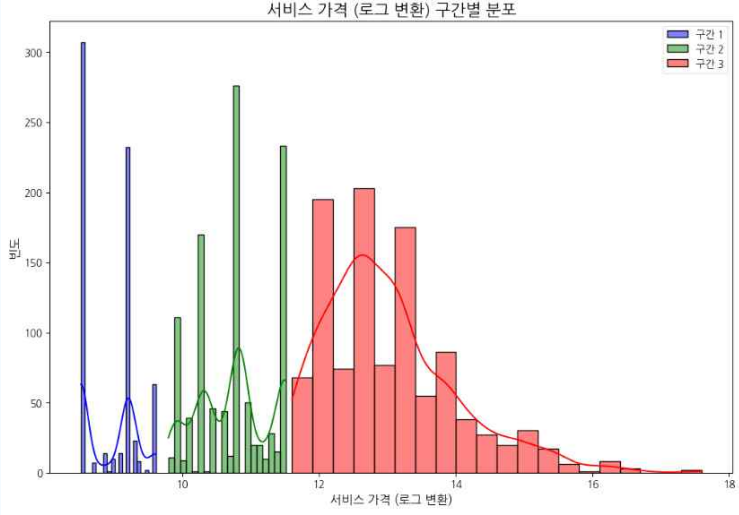
	group1	group2	meandiff	p-adj	lower	upper	reject
0	구간 1	구간 2	0.9694	0.0046	0.2498	1.6889	True
1	구간 1	구간 3	0.4931	0.2845	-0.2707	1.2568	False
2	구간 2	구간 3	-0.4763	0.2037	-1.1316	0.1790	False

거래 취소율에 대해 가격 구간 간에 유의미한 차이가 있습니다. (p-value = 0.005981888747067292)

P-value 0.006 < 0.05로 **거래 취소율과 구간별 가격 간 유의미한 관계가 있음을 확인**

구간1과 구간 2의 연관성은 있지만, (구간2와 3), (구간 1,3)의 연관성은 없음을 확인

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무

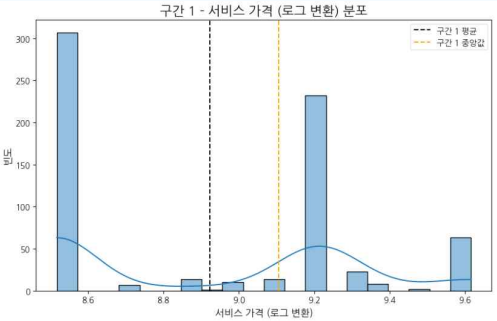


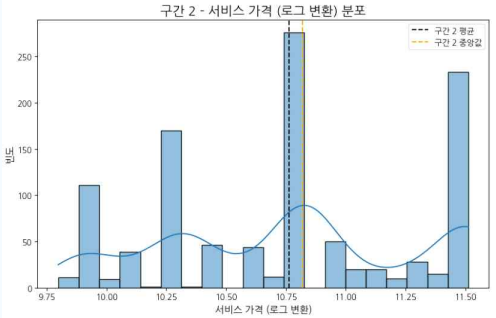
기준점1 = 15946.74 (서비스 가격)

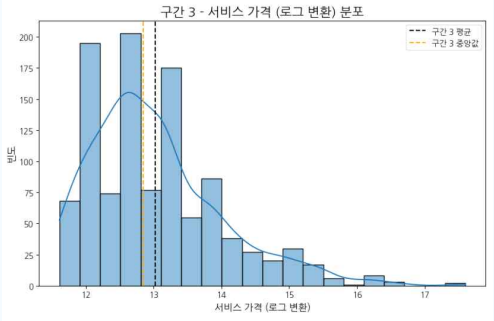
기준점2 = 105259.20 (서비스 가격)

기준점2의 구간 값을 새롭게 조정하여 추가적인 연관성을 분석하고자 함

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무







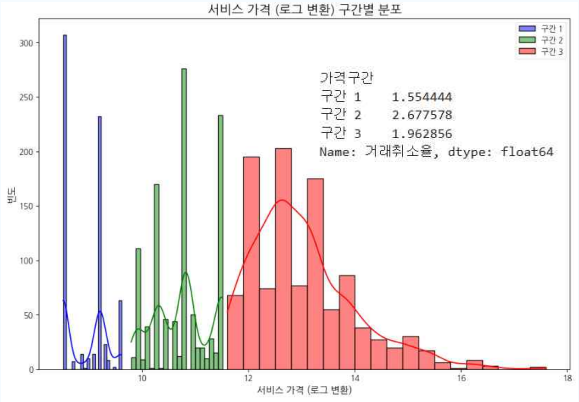
구간3을 재조정하여 각 서비스 가격의 분포를 다시 나타냄

구간 1 : 5000 ~ 15946.74

구간 2 : 15946.75 ~ 105259.19

구간 3 : 105259.20 ~ 44000000

3. (서비스) 가격에 따라(범주화) / (서비스) 거래 취소율에 차이가 있을 것이다. + 대분류 유무



ANOVA 분석 결과:

	sum_sq	df	F	PR(>F)
가격구간	584.277588	2.0	7.033309	0.000897
Residual	118752.753841	2859.0	NaN	NaN

Tukey HSD 사후 분석 결과:

	group1	group2	meandiff	p-adj	lower	upper	reject
0	구간 1	구간 2	1.1231	0.0011	0.3857	1.8605	True
1	구간 1	구간 3	0.4084	0.3974	-0.3304	1.1473	False
2	구간 2	구간 3	-0.7147	0.0261	-1.3619	-0.0675	True

거래 취소율에 대해 가격 구간 간에 유의미한 차이가 있습니다. (p-value = 0.0008973512510187171)

P-value 0.0009 < 0.05로 **거래 취소율과 구간별 가격 간 유의미한 관계가 있음을 확인**

구간1과 구간2, 구간2와 구간3의 연관성은 있음으로 파악되나, 구간1,3의 연관성은 없음

따라서, 중간 가격일수록 취소율이 낮고 저가나 고가일 경우 취소율이 높다는 결론을 도출함

4. 서비스의 특성이 판매 서비스 수에 미치는 영향 (상관분석)

가설: 서비스의 특성이 판매 서비스 수 사이의 관계

변수: 판매 서비스 수, 프리미엄 서비스 가입 여부/신속 알람 서비스/서비스 평점/서비스 가격/대분류

	상관계수	p-value
판매 서비스 수	1.0000	0.0000
총 판매 금액	0.5609	0.0000
신속 알람 서비스 사용여부	0.3660	0.0000
프리미엄 서비스 가입여부	0.3430	0.0000
이용자수	0.2395	0.0000
대분류_커머스	0.2048	0.0000
대분류_데이터	0.1496	0.0000
평점	0.0679	0.0000
대분류_게임	0.0555	0.0000
대분류_기타	0.0026	0.1324
대분류_기획	-0.0007	0.6775
거래취소율	-0.0316	0.0000
대분류_인공지능	-0.0350	0.0000
대분류_디자인	-0.0527	0.0000
대분류_홈페이지	-0.0761	0.0000
대분류_프로그램	-0.0790	0.0000
대분류_프로그래밍	-0.0953	0.0000
대분류_모바일	-0.1114	0.0000
서비스가격	-0.1518	0.0000
서비스번호	-0.3405	0.0000

$\alpha = 0.05 > p - value = 0$ => 유의미한 상관관계가 존재한다

4. 서비스의 특성이 판매 서비스 수에 미치는 영향 (상관분석)

```
변수 '대분류_디자인' 제거 (p-value=0.9307)
변수 '거래취소율' 제거 (p-value=0.2631)
변수 '대분류_기획' 제거 (p-value=0.1624)
변수 '대분류_프로그램' 제거 (p-value=0.2363)
                            OLS Regression Results
==============================================================================
Dep. Variable:                판매 서비스 수   R-squared:                       0.502
Model:                            OLS   Adj. R-squared:                  0.502
Method:                 Least Squares   F-statistic:                 2.413e+04
Date:                Fri, 28 Mar 2025   Prob (F-statistic):               0.00
Time:                        10:17:55   Log-Likelihood:            -2.6194e+06
No. Observations:              335782   AIC:                         5.239e+06
Df Residuals:                  335767   BIC:                         5.239e+06
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	-275.2729	7.146	-38.519	0.000	-289.280	-261.266
이용자수	1.0218	0.012	87.845	0.000	0.999	1.045
프리미엄 서비스 가입여부	-184.2636	3.119	-59.083	0.000	-190.376	-178.151
신속 알람 서비스 사용여부	309.9503	2.342	132.371	0.000	305.361	314.540
평점	-3.0920	1.300	-2.379	0.017	-5.639	-0.544
서비스가격	-0.0005	4.16e-06	-116.715	0.000	-0.000	-0.000
총 판매 금액	2.723e-05	8.03e-08	339.269	0.000	2.71e-05	2.74e-05
대분류_게임	852.0527	11.473	74.266	0.000	829.566	874.539
대분류_기타	337.6718	4.330	77.993	0.000	329.186	346.158
대분류_데이터	549.1160	4.079	134.612	0.000	541.121	557.111
대분류_모바일	-156.1890	5.779	-27.025	0.000	-167.516	-144.862
대분류_인공지능	-58.7696	16.271	-3.612	0.000	-90.661	-26.878
대분류_커머스	760.8036	4.964	153.264	0.000	751.074	770.533
대분류_프로그래밍	187.6340	4.674	40.141	0.000	178.472	196.796
대분류_홈페이지	225.0676	3.718	60.538	0.000	217.781	232.354

Omnibus:	114347.297	Durbin-Watson:	0.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1393508.132
Skew:	1.292	Prob(JB):	0.00
Kurtosis:	12.640	Cond. No.	5.07e+08

회귀분석 중 단계 선택법 (후진 제거법) 사용하여 회귀모델의 성능을 저해하는 요소들을 제거 후 분석

(p-value =0.05)

R-squared: 0.502로 회귀식이 약 50%의 설명력을 가진다.

인사이트 도출

1. 추가 결제 옵션이 있는 서비스는 사용자 만족도가 상대적으로 높아 취소율이 낮을 가능성이 있음

사용자 만족도가 거래 취소율에 유의미한 영향을 미치는지 추가적인 분석이 필요하지만 데이터가 없어 불가능함

2. 수수료율이 높을수록 매출이 증가하는 경향이 있음

3. 중간 가격일수록 취소율이 낮고 저가나 고가일 경우 취소율이 높음

4. 서비스의 특성이 판매 서비스 수에 유의미한 영향이 있는 것으로 확인됨

(판매 서비스 수에 가장 많은 영향을 미치는 변수 : 서비스 中 게임, 커머스 / 신속 알람 서비스 여부)

L사의 비즈니스적 전략 제안

1. 서비스 가격에 대한 민감도 낮음 → 고객 경험 강화가 핵심

결제 시스템 및 UI 개편 등을 통해 **가격 외 가치를 제공하고 고객 만족도 향상 및 서비스 품질 개선**

중간 가격대의 서비스가 거래 취소율을 낮추고, 고객 이탈을 방지하는 데 유리

2. 가격 변동 및 추가 결제 금액이 거래 취소율에 미치는 영향 미미 → 타겟 마케팅과 맞춤형 혜택 제공이 핵심

고객 패턴 분석을 통한 맞춤형 서비스 제공

재구매 고객에게 할인, 포인트 등의 혜택을 제공해 재구매 유도

3. 수수료율이 높을수록 매출 증가 경향 → 수수료 개편을 통한 매출 증대 및 고객 만족도 향상

거래 유형에 맞춘 수수료 차등화 및 할인 방안 도입으로 고객 부담을 줄이고, 재구매 유도 및 신뢰도 증가

4. 서비스 특성과 판매 수 간 유의미한 상관관계 → 판매 전략 최적화

판매자의 신속 알람 서비스 기본화, 커머스 서비스의 특화된 마케팅(할인 쿠폰, 마일리지 적립 등)을 통한 고객 유치

iM뱅크

감사합니다