

# 온라인 IT교육 유입 요인 분석

프로젝트 진행 기간

2024.01.15 ~ 2024.01.19 15:00 (총 5일)

1조 팀원

심현지, 이신영, 이인철, 이효준, 조용재

# 요약

01

...

02

...

03

수강생 수에 영향을  
미치는 변수를  
의사결정나무 분석으로  
확인

Mann-Whitney U  
검정  
↓  
각 변수에서의  
구매 경향 확인

가격, 강의시간, 평점,  
난이도 측면에서  
구매 경향을 바탕으로  
적절한 마케팅 전략 제안

# 목차

2024

빅데이터 6기 크롤링 프로젝트 결과보고서

## 01

### 프로젝트 개요

업무 분류 체계(WBS)	4
목적 및 배경	5
데이터 분석 활용 시나리오	6
기술 및 기능	9
프로젝트 지원	10

## 02

### 데이터 분석

데이터 선정	11
데이터 분석 결과	13

## 03

### 분석 활용 전략

방향 제시	28
-------	----

# 1. 프로젝트 개요

- 프로젝트명: 온라인 IT교육 유입 요인 분석
- 프로젝트 진행 기간: 2024.01.15 ~ 2024.01.19 15:00(총 5일)
- 역할 분배(R&R, Role and Responsibility)

	책임 파트	상세 업무
심현지	데이터 분석, 결과 정리	다중회귀 분석, 상관관계 분석, 의사결정나무 분석, 최종 보고서(PPT) 제작
이신영	데이터 분석, 결과 정리, 발표	탐색적 데이터 분석, 최종 보고서(PPT) 제작, 발표
이인철	데이터 수집	크롤링, 데이터 분석 서포트
이효준	데이터 분석	탐색적 데이터 분석
조용재	데이터 수집, 데이터 분석	크롤링, 데이터 분석 서포트

# 1. 프로젝트 개요

## 업무 분류 체계(WBS)

진행 단계	세부 단계	활동 설명	담당자	프로젝트 일정 기간				
				1Day	2Day	3Day	4Day	5Day
착수	프로젝트 착수	제안요약서 확인	전체					
계획	프로젝트 계획	착수 보고서 작성	전체					
실행	데이터 크롤링	웹사이트 구조 파악	이인철, 조용재					
		모듈화 생성						
		동작 테스트 및 디버깅						
	탐색적 데이터 분석	데이터 구조 파악 (데이터의 크기, 변수의 수, 데이터 타입 확인)	이신영, 이효준					
		결측치 확인 및 결측치 처리	이신영, 이효준					
		기초 통계 분석 (평균, 중앙값, 표준편차)	이신영, 이효준					
		시각화 (변수의 분포 시각화)	이신영, 이효준					
		데이터 전처리	이신영, 이효준					
		피어슨 상관관계 분석	심현지					
		상관관계 분석 시각화	심현지					

회귀분석	회귀분석	모델 경향성 확인 *피어슨 상관관계 분석 이용	심현지					
		모델 적합성 확인 (정규성, 등분산성 확인)	심현지					
		회귀계수 계산 및 유의성 확인 (독립변수 간 다중공선성, 회귀계수 유의성, 독립변수 선택 및 해석)	심현지					
		최종 회귀 모델 선정	심현지					
	의사결정 나무 분석	변수 선택	심현지					
		의사결정나무 생성	심현지					
		의사결정나무	심현지					
		평가						
		의사결정나무 주요 변수 시각화	심현지					
	종료	최종 보고서 작성 및 제출	심현지, 이신영					
		발표	이신영					

# 1. 프로젝트 개요

목적 및 배경

분석 대상: 인프런(InFlearn)

클라이언트의 니즈인 '수강자 수 유입에 영향을 미칠 수 있는 데이터' 수집에 적합

	패스트캠퍼스	인프런	스파르타코딩클럽
강의명	O	O	O
강의 카테고리	O	O	O
총 강의 시간	O	O	O
평점	X	O	X
가격	O	O	O
무료/유료	X	O	X
수강생 인원	X	O	X
난이도	X	O	X

[온라인 IT교육 3사 비교]

# 1. 프로젝트 개요

데이터 분석 활용 시나리오

비즈니스 문제 정의

신규 IT 교육 서비스 전략 수립 및 활용을 위한 인프라 정보 수집 및 데이터 분석 진행

	데이터 타입	데이터 예시	수집 방법
강의명	string	김영한의 실전 자바- 기본편	셀레니움을 이용한 웹 크롤링
강의 카테고리	string	개발·프로그래밍	
총 강의 시간	string	16시간 51분	
평점	float	5.0	
가격	int	44,000	
무료/유료	string	무료/유료	
수강생 인원	int	5,050	
난이도	string	입문/초급/중급자	

[데이터 수집 방법 및 목록]

# 1. 프로젝트 개요

## 분석 기법

탐색적 데이터 분석	<ul style="list-style-type: none"> <li>수집한 데이터를 다양한 각도에서 관찰·이해하기 위해 분석 전 그래프를 그리거나 통계적인 방법으로 데이터를 직관적으로 바라보는 방법</li> <li>데이터의 잠재적인 문제와 다양한 패턴 발견 가능</li> <li>적절한 통계 도구, 자료 수집 제시 가능</li> </ul>
상관관계 분석	<ul style="list-style-type: none"> <li>두 변수 간의 선형적 관계를 분석하는 기법으로 상관계수를 이용하여 측정하는 방법론</li> <li>피어슨 상관계수: 상관분석에서 기본적으로 사용되는 상관계수로, 연속형 변수의 상관관계를 측정</li> <li>절대값 1.0 ~ 0.7 : 매우 높은 음/양의 상관관계</li> <li>절대값 0.7 ~ 0.3 : 높은 음/양의 상관관계</li> <li>절대값 0.3 ~ 0 : 낮은 음/양의 상관관계</li> <li>0 : 상관관계 없음</li> </ul>



# 1. 프로젝트 개요

## 분석 기법

### 다중선형회귀 분석

- 여러 개의 독립변수가 종속변수에 미치는 영향을 분석하는 통계적인 방법
- 모델은 주어진 데이터를 기반으로 회귀 계수를 추정하여, 이를 사용해 새로운 입력 값에 대한 종속 변수의 값을 예측
- 다중 선형 회귀식:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$

### 의사결정나무 분석

- 일련의 분류 규칙을 통해 데이터를 분류(classification) 혹은 회귀(regression)할 수 있는 지도 학습 모델 중 하나
- 시각화에 효과적이기 때문에 알고리즘이 쉽고 직관적임
- 통계모델에 요구되는 가정에 자유로움 (비모수적 모델)

# 1. 프로젝트 개요

통계적 가설 설정

귀무가설  $H_0 : p = 0$

평점, 강의 시간, 난이도, 가격, 무/유료 변수가 인프런 IT 교육 수강자 수에 영향을 주지 않는다.

대립가설  $H_1 : p \neq 0$

평점, 강의 시간, 난이도, 가격, 무/유료 변수가 인프런 IT 교육 수강자 수에 영향을 준다.

# 1. 프로젝트 개요

프로젝트 지원

## 제약사항

유/무료, 가격 두 변수로 인한 다중공선성 문제 예상

## 대응방안

- 각 입력 변수를 제거/추가하면서 회귀계수의 변동 정도 파악
- 상관계수가 높은 독립변수 중 하나 혹은 일부를 회귀모형에서 제거
- 변수를 변형시키거나 새로운 관측치 이용
- 독립변수에 의존하는 변수 제거(VIF)

## 하자보수 목적

사용자에게 나오는 오류 수집, 문제의 신속한 해결

## 수행 날짜

2023.01.17 (수)

## 수행 방법

- 프로젝트를 수행하는 5인의 파이썬 환경에서 프로그램 동작 여부 확인
- 프로그램이 동작하지 않을 시, XPath 변경 여부 확인 작업 수행 및 코드 수정

## 2. 데이터 분석

데이터 선정

	데이터 수	데이터 타입	데이터 예시	수집 방법
강의명	1351	object	김영한의 실전 자바- 기본편	셀레니움을 이용한 웹 크롤링
강의 카테고리		string	개발·프로그래밍	
총 강의 시간		float64	16시간 51분	
평점		float64	5.0	
가격		int32	44,000	
무료/유료		int64	무료/유료	
수강생수		int32	5,050	
난이도		string	입문/초급/중급자	

[데이터 수집 방법 및 목록]

## 2. 데이터 분석 분석 프로세스



...



...



...



### 탐색적 데이터 분석

- 데이터 구조 파악
- 결측치, 이상치 처리
- 기초 통계 분석
- 시각화

### 상관관계 분석

- 피어슨 상관관계
- 시각화

### 다중선형회귀 분석

- 모델 경향성 확인
- 모델 적합성 확인

### 의사결정나무 분석

- 의사결정나무 학습/훈련
- 의사결정나무  
주요 변수 시각화

## 2. 데이터 분석

탐색적 데이터 분석

### 데이터 전처리 결과

	강좌명	수강생수	가격	평점	수강평수	난이도	강의시간_분	무료/유료
	스프링 핵심 원리 - 기본편	30486	61600	5.0	326	초급자	725	유료
	김영한의 실전 자바 - 기본편	5147	30800	5.0	205	초급자	1011	유료
	스프링 부트 - 핵심 원리와 활용	8595	69300	5.0	203	초급자	945	유료
	모든 개발자를 위한 HTTP 웹 기본 지식	27573	30800	5.0	832	초급자	340	유료
	[2024 NEW] 개발자를 위한 쉬운 도커	170	61600	5.0	3	초급자	594	유료

## 2. 데이터 분석

### 탐색적 데이터 분석

요인분석을 위한 데이터 전처리

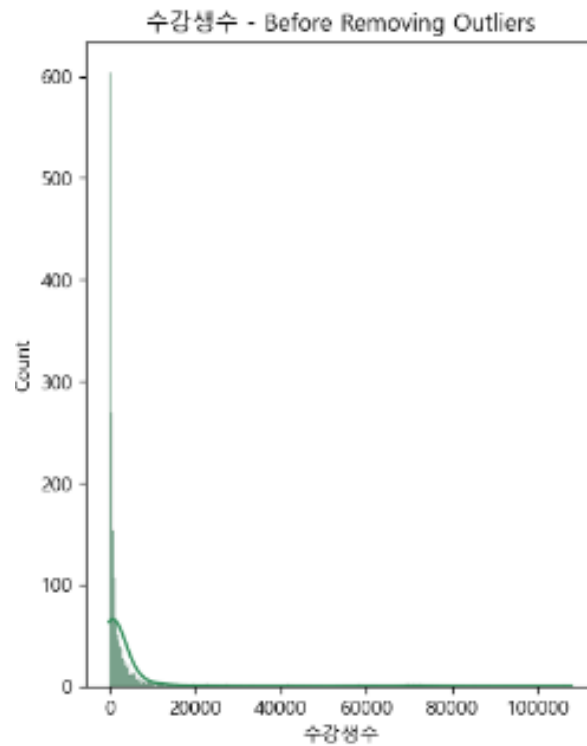
난이도	'입문자/초급자/중급자' → 각각의 해당 여부 더미변수로 변환
무료/유료	'유료/무료' → '유료 : 1, 무료 : 0'으로 더미변수 변환

강좌명	수강생 수	가격	평점	수강 평수	강의 게시 일	강의 시간 분	평가지 수	강의 시간당 가격	난이도_입문자	난이도_중급자	난이도_초급자	유무료
스프링 핵심 원리 - 기본편	30486	61600	5.0	326	2020-09-21	725	1630.0	5097	False	False	True	1
김영한의 실전 자바 - 기본편	5147	30800	5.0	205	2023-11-28	1011	1025.0	1827	False	False	True	1

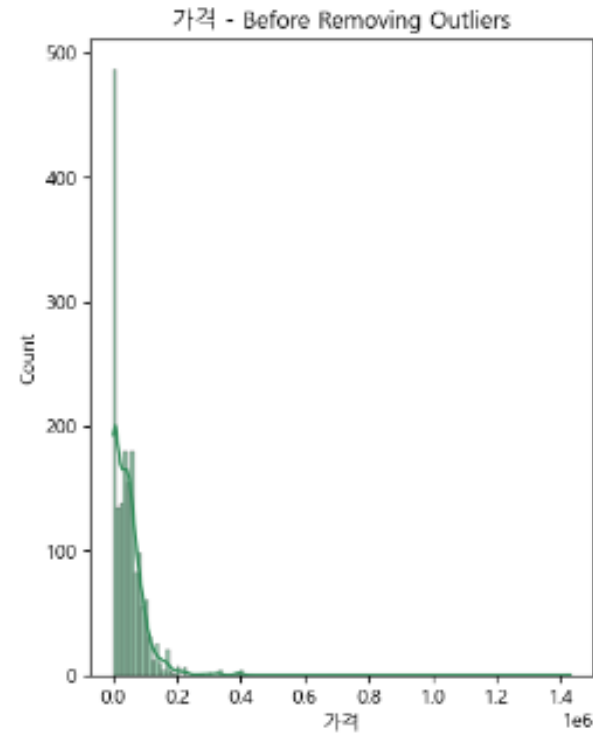
## 2. 데이터 분석

탐색적 데이터 분석

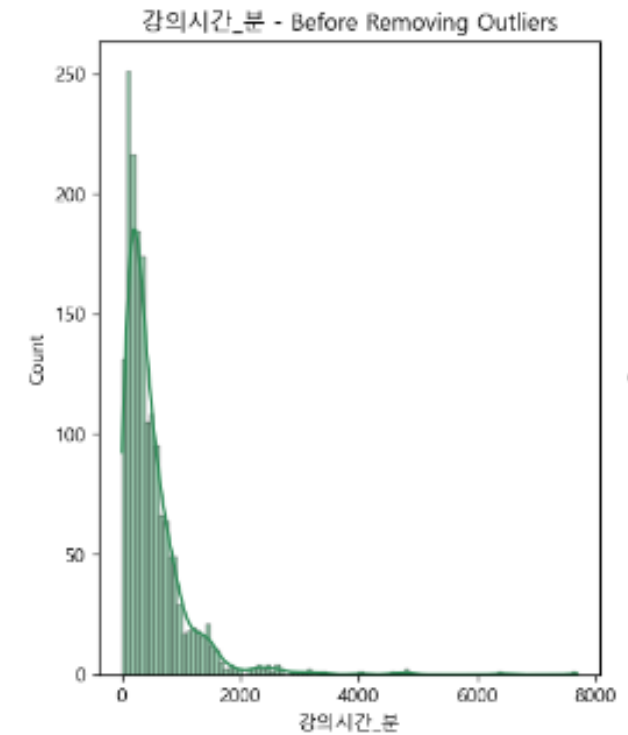
### 데이터 분포



수강생수



가격



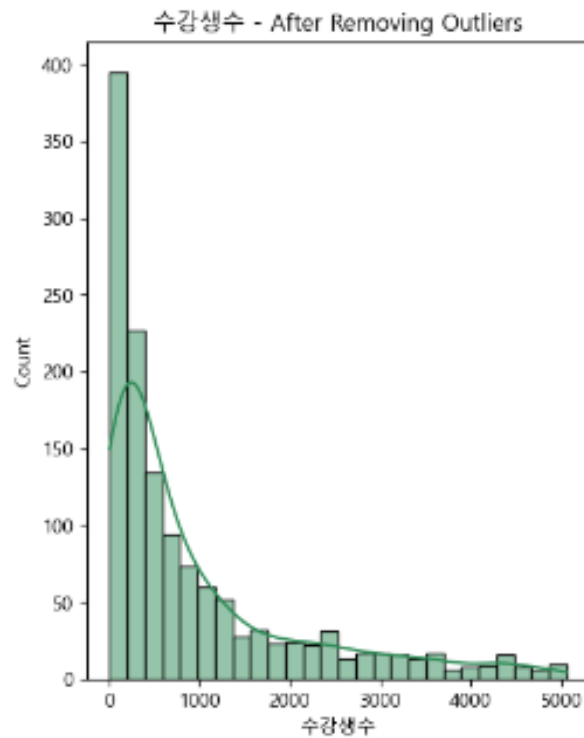
강의시간



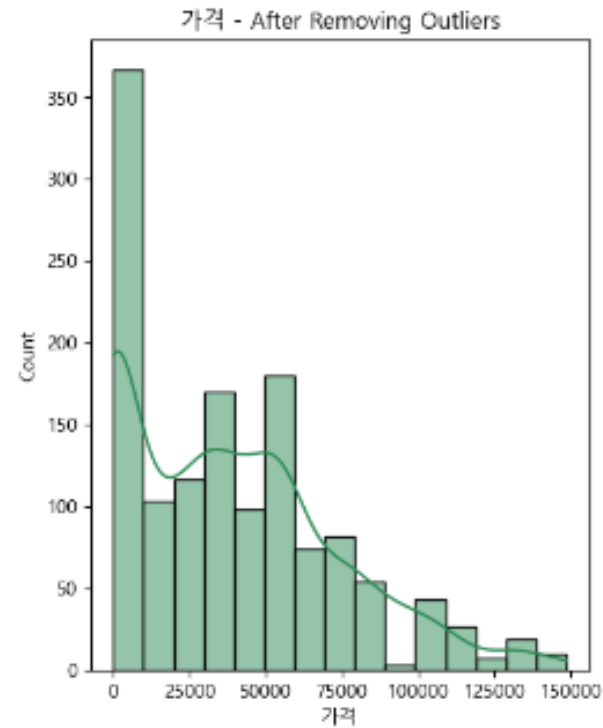
## 2. 데이터 분석 탐색적 데이터 분석

데이터 분포

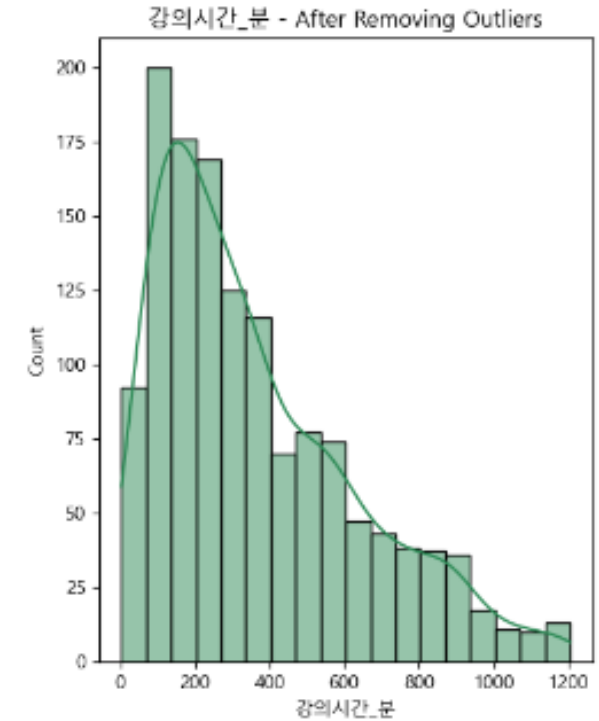
세 변수에 대한 IQR 기반 이상치 제거 결과



수강생수



가격

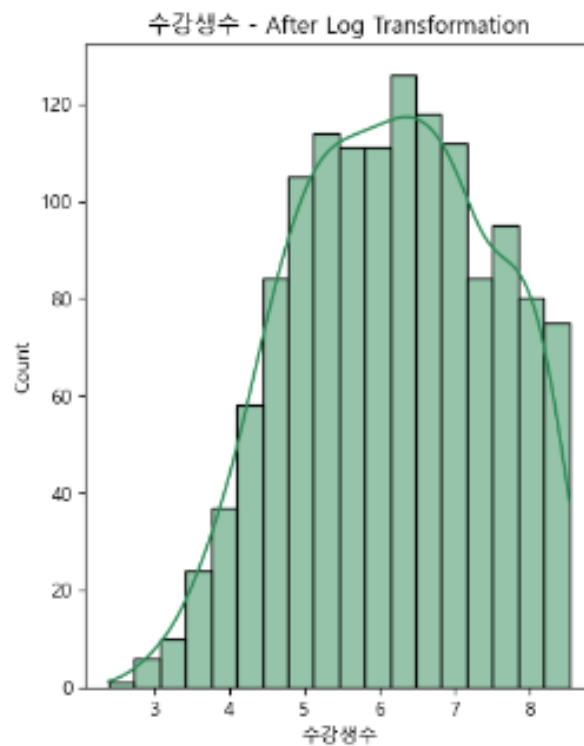


강의시간

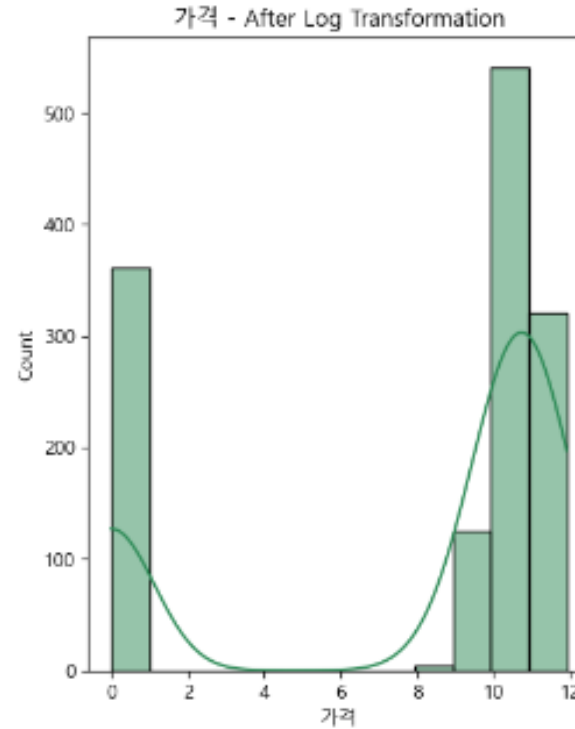
## 2. 데이터 분석 탐색적 데이터 분석

데이터 분포

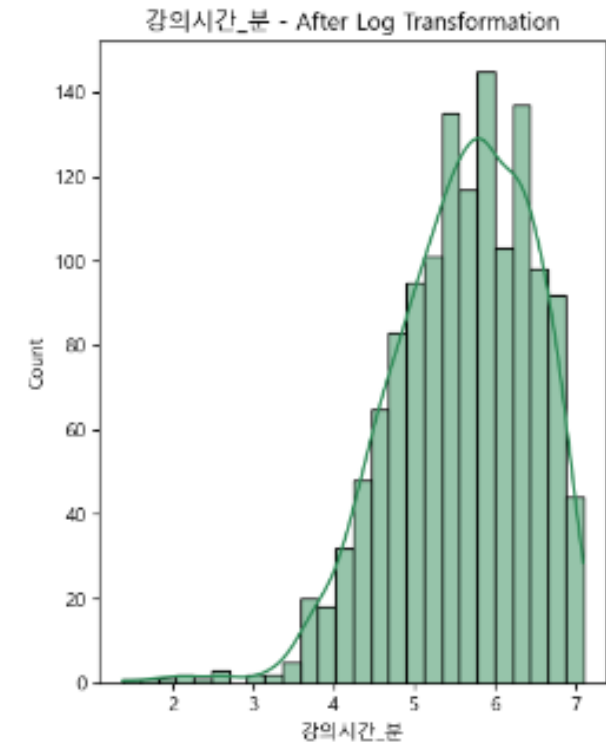
세 변수에 대한 로그 스케일링 결과



수강생수



가격



강의시간

## 2. 데이터 분석

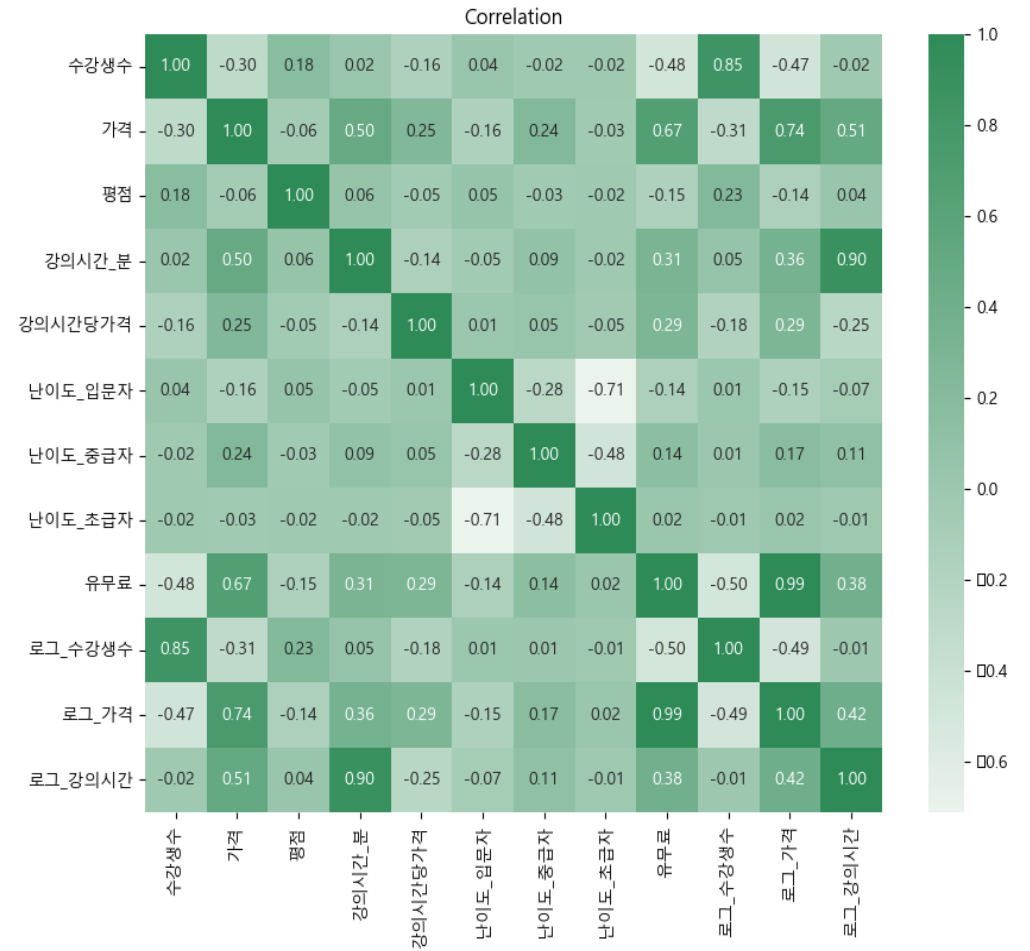
### 탐색적 데이터 분석

#### 기초 통계

	수강생수	가격	평점	강의시간_분	강의시간당가격	난이도_입문자	난이도_중급자	난이도_초급자	유무료	로그_수강생수	로그_가격	로그_강의시간
count	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000	1351.000000
mean	976.353812	38158.985936	4.714286	365.120651	8065.102887	0.291636	0.156181	0.552184	0.732791	6.154129	7.825414	5.592572
std	1159.038139	34443.637737	0.378940	268.006934	17048.432565	0.454684	0.363161	0.497454	0.442666	1.300267	4.758753	0.863951
min	10.000000	0.000000	1.300000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.397895	0.000000	1.386294
25%	173.000000	0.000000	4.600000	151.500000	0.000000	0.000000	0.000000	0.000000	0.000000	5.159055	0.000000	5.027159
50%	480.000000	33000.000000	4.800000	298.000000	5799.000000	0.000000	0.000000	1.000000	1.000000	6.175867	10.404293	5.700444
75%	1274.000000	55000.000000	5.000000	527.000000	10530.000000	1.000000	0.000000	1.000000	1.000000	7.150701	10.915107	6.269096
max	5064.000000	148500.000000	5.000000	1204.000000	396000.000000	1.000000	1.000000	1.000000	1.000000	8.530109	11.908347	7.094235

## 2. 데이터 분석 상관계수 분석

- 독립변수 중 **가격 변수만 -0.3** 으로 약한 음의 상관계수 존재
- 그 외는 **0.3** 미만으로 **상관계수가 존재하지 않음**
- \* 로그 변환한 변수는 로그 변환 전 본 변수와의 상관계수는 무시



## 2. 데이터 분석 정규성, 등분산성 검정

```
print(shapiro(df1['로그_수강생수']))
print(anderson(df1['로그_수강생수']))
print(levene(df1['로그_수강생수'], df1['로그_가격']))

ShapiroResult(statistic=0.9822836518287659, pvalue=8.121748933109796e-12)
AndersonResult(statistic=4.930499091092997, critical_values=array([0.574, 0.65
4, 0.785, 0.915, 1.089]), significance_level=array([15. , 10. , 5. , 2.5,
1. ]))
LeveneResult(statistic=308.562503459462, pvalue=1.7207205812348635e-65)
```

### Shapiro 검정 결과

p-value 값이 0.05 보다 작으므로  
데이터는 정규분포를 따르지 않는다.

### Levene 검정 결과

p-value 값이 0.05 보다 작으므로  
데이터는 등분산성을 충족하지 않는다.

## 2. 데이터 분석 다중선형회귀 분석

- 결정 계수: 모델이 데이터의 약 **31.7%**의 변동 설명
- p-value: 유/무료를 제외한 나머지 변수는 0.05 이하로 모델 전체가 **통계적으로 유의미**

OLS Regression Results						
=====						
Dep. Variable:	로그_수강생수		R-squared:	0.317		
Model:	OLS		Adj. R-squared:	0.314		
Method:	Least Squares		F-statistic:	103.8		
Date:	Thu, 18 Jan 2024		Prob (F-statistic):	1.86e-107		
Time:	15:50:00		Log-Likelihood:	-2014.1		
No. Observations:	1351		AIC:	4042.		
Df Residuals:	1344		BIC:	4079.		
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	2.4059	0.315	7.633	0.000	.788	3.024
로그_가격	-0.1258	0.060	-2.112	0.035	-0.243	-0.009
평점	0.5088	0.079	6.449	0.000	0.354	0.664
로그_강의시간	0.3260	0.040	8.127	0.000	0.247	0.405
난이도_입문자	0.6276	0.116	5.414	0.000	0.400	0.855
난이도_중급자	1.0099	0.120	8.441	0.000	0.775	1.245
난이도_초급자	0.7684	0.111	6.907	0.000	0.550	0.987
유무료	-0.3470	0.627	-0.553	0.580	-1.578	0.884
=====						
Omnibus:	2.444	Durbin-Watson:	1.622			
Prob(Omnibus):	0.295	Jarque-Bera (JB):	2.339			
Skew:	-0.053	Prob(JB):	0.310			
Kurtosis:	2.825	Cond. No.	1.28e+17			
=====						

## 2. 데이터 분석 다중선형회귀 분석

정규성 / 등분산성 / 선형

자유로운

- 결정 계수: 모델이 데이터의 약 31.7%의 변수 설명
- p-value: 유/무료를 제외한 설정한 나머지 변수는 0.05 이하로 모델 전체가 통계적으로 유의미

의사결정나무 모델 도입!

성 / 선형에

로운

모델 도입!

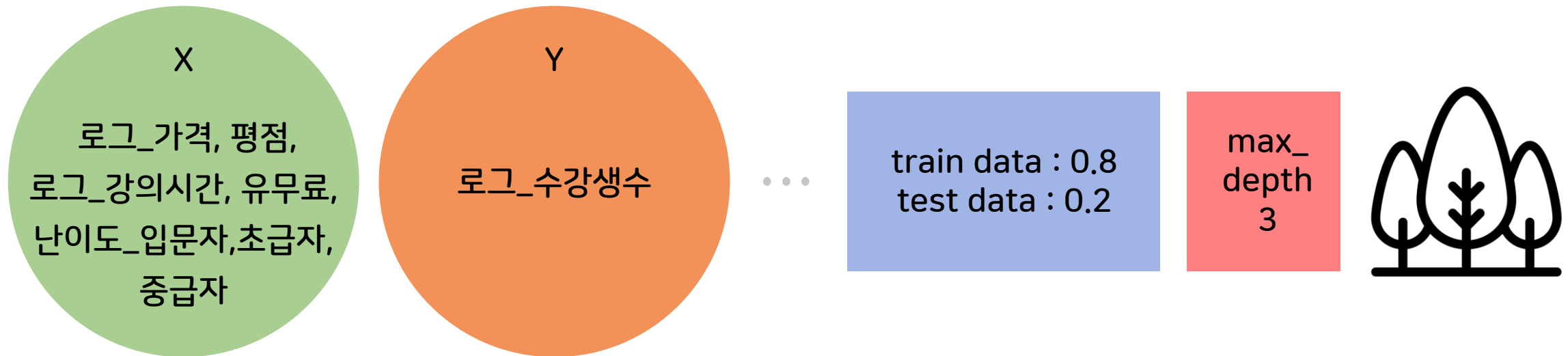
OLS Regression Results

Dep. Variable:	로그_수강생수	R-squared:	0.317
Model:	OLS	Adj. R-squared:	0.314
Method:	Least Squares	F-statistic:	103.8
Date:	Thu, 17 May 2024	Prob(F-statistic):	1.86e-107
Time:	10:55:00	Log Likelihood:	-2014.1
No. Observations:	351	AIC:	4042.
Df Residuals:	1344	BIC:	4079.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.4059	0.315	7.633	0.000	1.788	3.024
로그_가격	-0.1258	0.060	-2.112	0.035	-0.243	-0.009
평점	0.5088	0.070	6.449	0.000	0.354	0.664
로그_간이료	0.0260	0.000	8.127	0.000	0.0247	0.0405
난이도_기분	0.0016	0.000	5.414	0.000	0.0004	0.0026
난이도_중급자	1.0099	0.120	8.441	0.000	0.775	1.245
난이도_초급자	0.7684	0.111	6.907	0.000	0.550	0.987
유무료	-0.3470	0.627	-0.553	0.580	-1.578	0.884

Omnibus:	2.444	Durbin-Watson:	1.622
Prob(Omnibus):	0.295	Jarque-Bera (JB):	2.339
Skew:	-0.053	Prob(JB):	0.310
Kurtosis:	2.825	Cond. No.	1.28e+17

## 2. 데이터 분석 의사결정나무 분석





## 2. 데이터 분석 의사결정나무 분석

```
# 훈련 데이터 예측
y_train_pred = tree_model.predict(X_train)

# 훈련 데이터 평가 지표 출력
train_mse = mean_squared_error(y_train, y_train_pred)
print("Training Mean Squared Error:", train_mse)

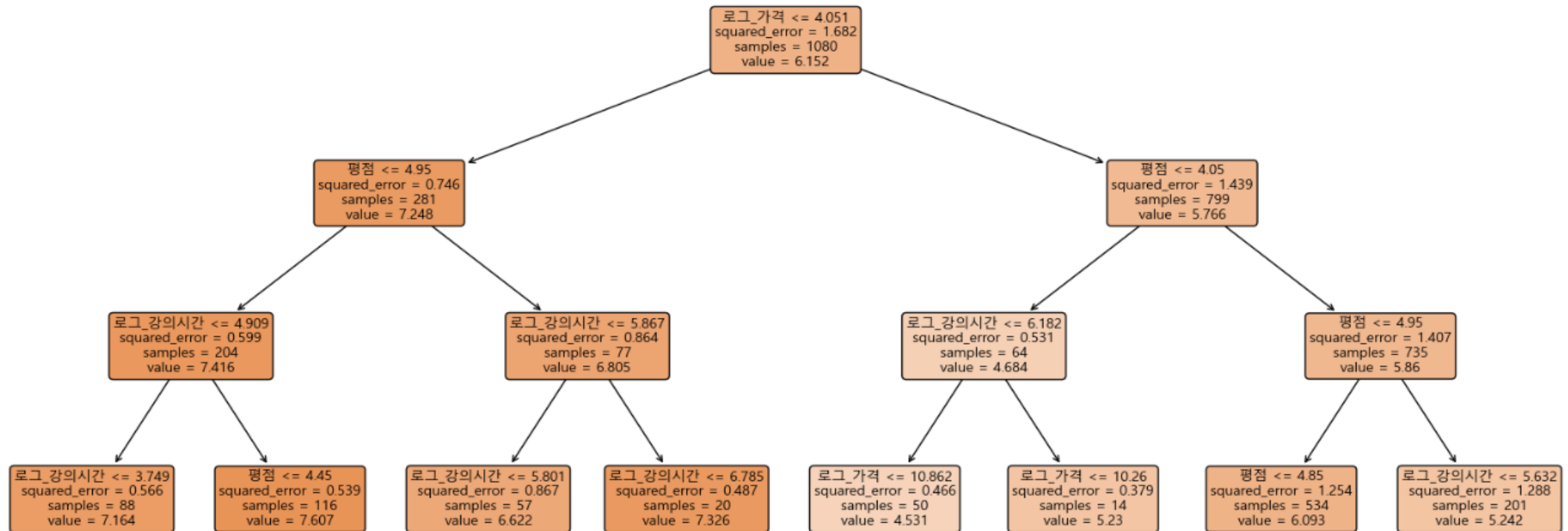
# 테스트 데이터 평가 지표 출력
test_mse = mean_squared_error(y_test, y_pred)
print("Test Mean Squared Error:", test_mse)

Training Mean Squared Error: 1.0451317852748327
Test Mean Squared Error: 1.1018105541476864
```

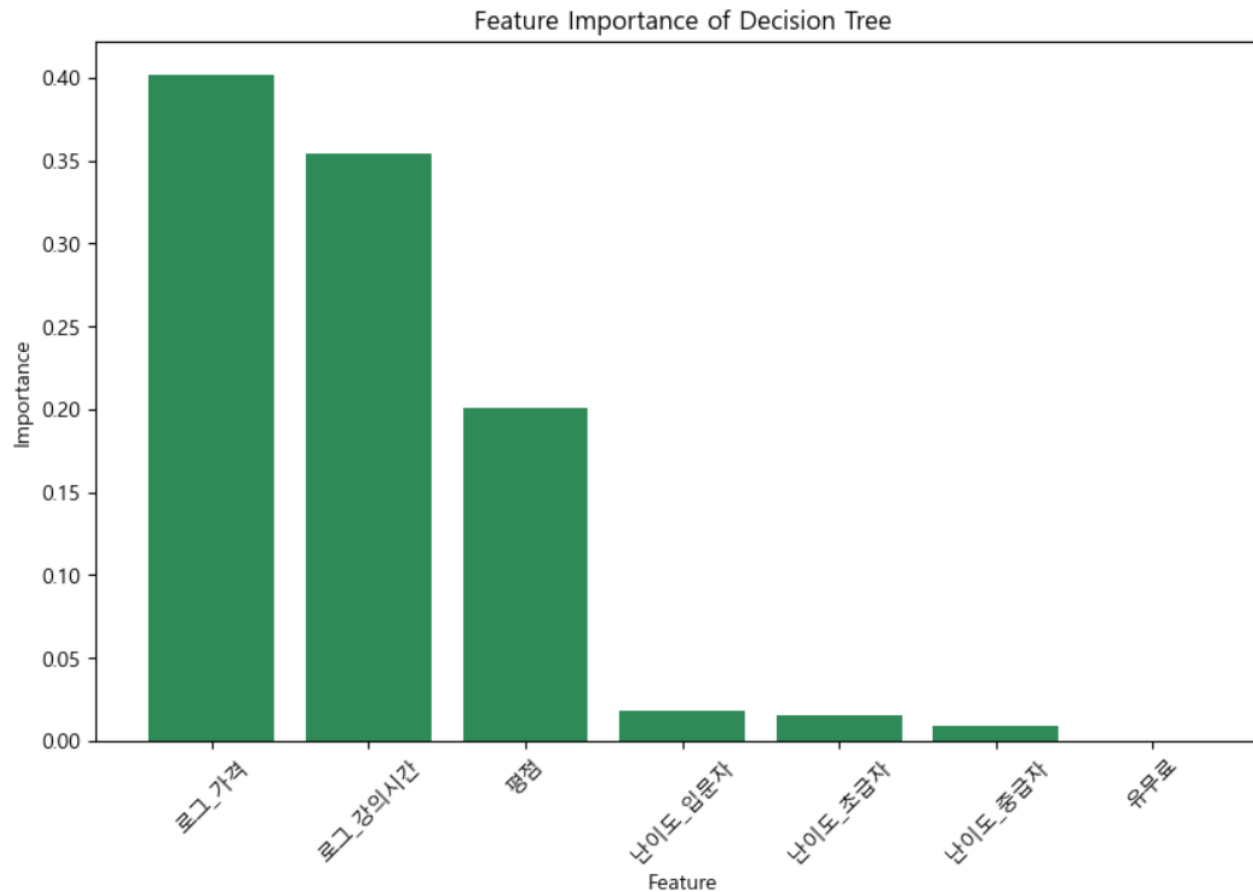
### MSE 결과

Train MSE : 약 1.04의 오차항을 가짐  
Test MSE : 약 1.10의 오차항을 가짐

## 2. 데이터 분석 의사결정나무 분석



## 2. 데이터 분석 의사결정나무 분석

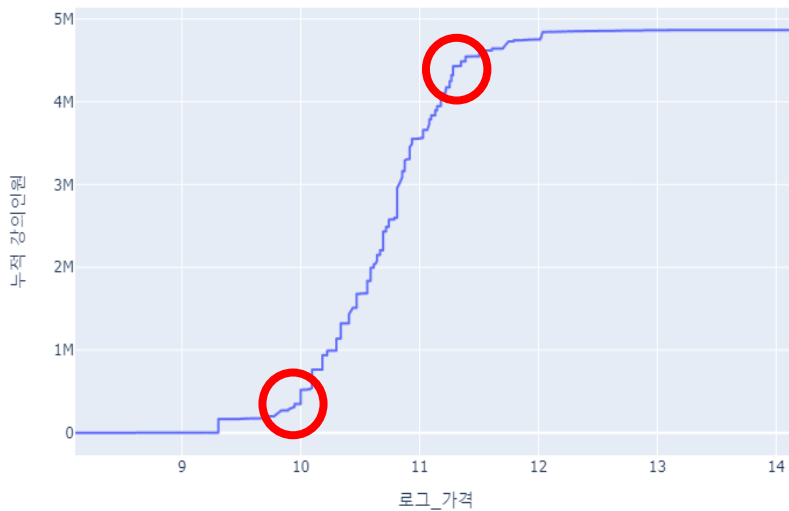


주요 Feature

수강생수에 가장 큰 요인을 주는 변수  
로그\_가격, 로그\_강의시간, 평점

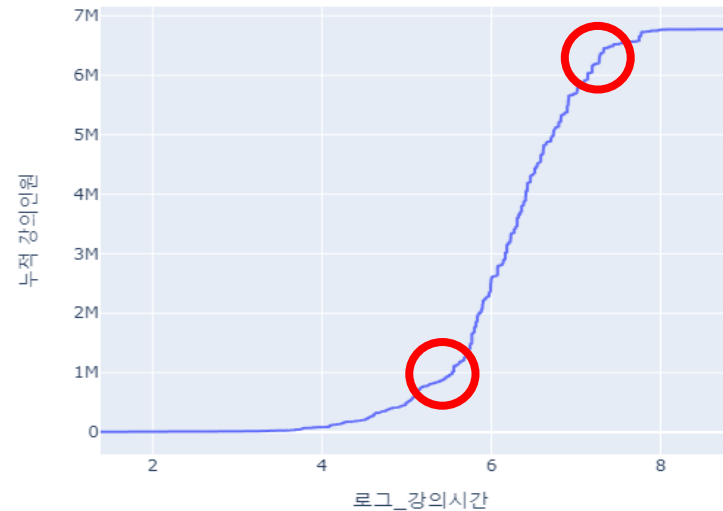
## 2. 데이터 분석 Mann-Whitney U 검정

가격별 누적 강의인원



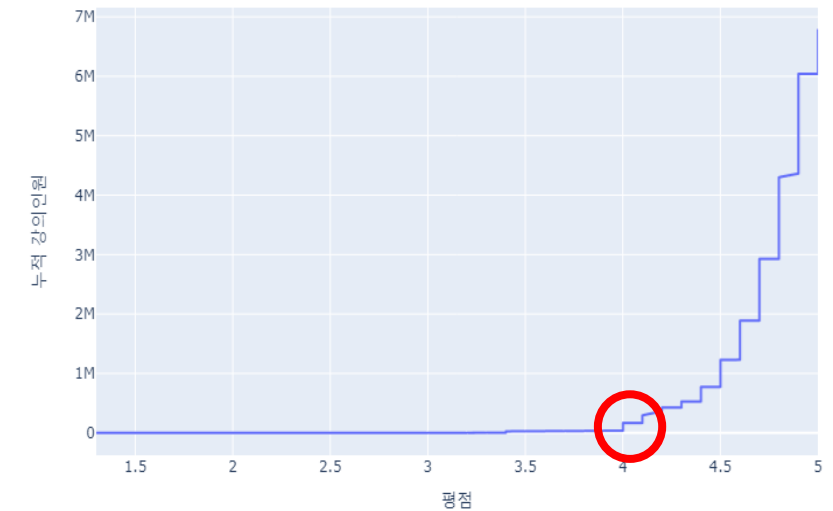
수강생의 강의 구매 경향이 나타난  
가격 구간  
: 약 22,026원 ~ 84,120원

강의시간별 누적 강의인원



수강생의 강의 구매 경향이 나타난  
강의시간 구간  
: 약 251분 ~ 1490분

평점별 누적 강의인원



수강생의 강의 구매 경향이 나타난  
평점 구간  
: 약 4점 ~ 5점

### 3. 분석 활용 전략

방향 제시

#### 가격 관련

- 가격 조정 및 할인 프로모션: 가격 범위 약 22,000원 ~ 84,000원
- 할인 프로모션: 약 84,000원 이상 강의의 경우 게시 초기 할인 프로모션 진행

#### 강의시간 관련

- 강의 시간 조정: 강의 시간 범위 약 251분(4시간 11분) ~ 1490분(24시간 50분)
- 강의 시간이 지나치게 짧을 경우 → 여러 강의 묶음 진행하는 방식
- 강의 시간이 지나치게 길 경우 → 여러 강의로 나누어 진행하는 방식

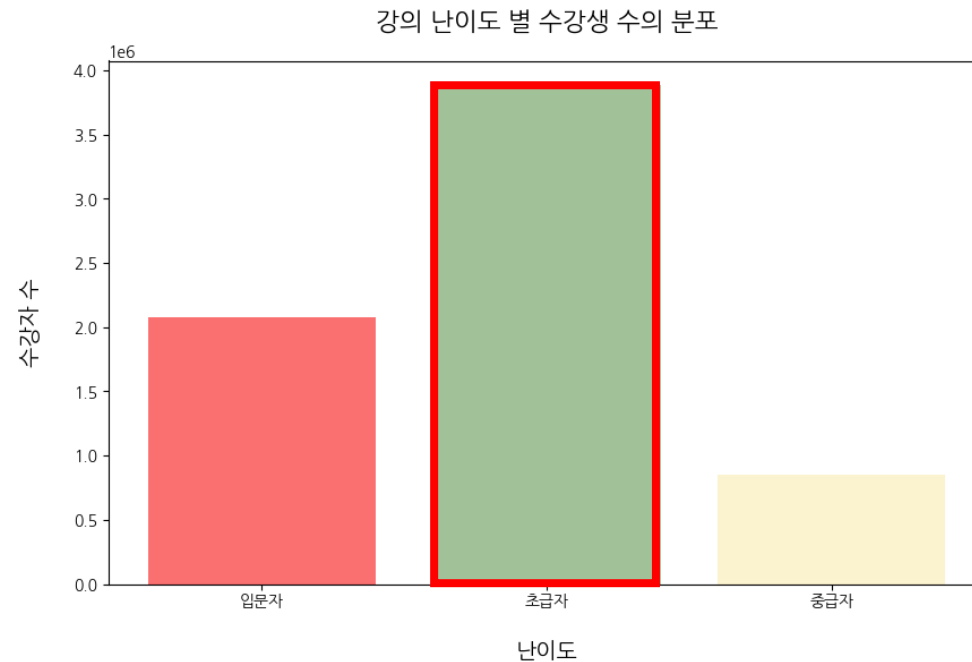
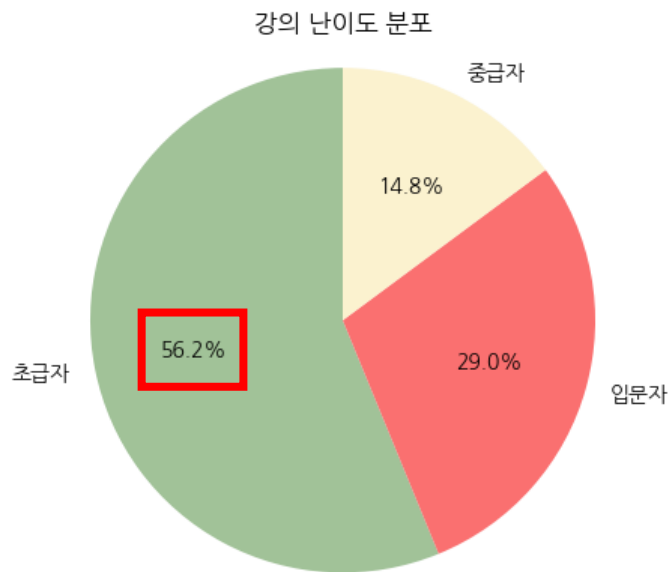
#### 평점 관련

- 4점 이하 강의평의 강의는 개별 피드백 제공
- 강의 게시 일정 기간 지났지만 평점 낮거나 없는 경우 프로모션 진행 또는 강사 개별 피드백
- 수강평 이벤트 진행: 수강 강의에 높은 평점을 주면 추가 강의 및 부가 서비스를 무료로 제공하는 등의 보상 및 리워드 프로그램을 도입

### 3. 분석 활용 전략 방향 제시

#### 난이도 관련

- 맞춤형 교육 계획 수립: 수강생들이 선호하는 난이도의 강의를 늘리는 맞춤형 교육 계획 수립
- 마케팅 및 광고 전략 구성: 초급자 대상 강의 홍보 및 할인 혜택 부여
- 강의 개발 방향성 도출: 초급자 대상 강의 개발 또는 기존 강의 업데이트



인과관계가  
불명확하므로  
관련 추가 분석  
진행 가능

# 코드 및 참고문헌

## 코드

[https://github.com/lpfe/20240115\\_20240119](https://github.com/lpfe/20240115_20240119)

## 참고문헌

서울대학교 AI연구원. (2022년8월30일). 탐색적 데이터 분석. 네이버 포스트.

<https://post.naver.com/viewer/postView.naver?volumeNo=34396755&memberNo=58684429&navigationType=push>

서울대학교 AI연구원. (2022년8월30일). 다중 회귀 분석. 네이버 포스트.

<https://post.naver.com/viewer/postView.naver?volumeNo=34397050&memberNo=58684429&navigationType=push>

Yan-yan SONG. (2015, Apr 25). Decision tree methods: applications for classification and prediction. PubMed Central. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/>

김진석. (2022년6월7일). 데이터과학.

[http://bigdata.dongguk.ac.kr/lectures/datascience/\\_book/%EC%9D%98%EC%82%AC%EA%B2%B0%EC%A0%95%EB%82%98%EB%AC%B4tree-model.html](http://bigdata.dongguk.ac.kr/lectures/datascience/_book/%EC%9D%98%EC%82%AC%EA%B2%B0%EC%A0%95%EB%82%98%EB%AC%B4tree-model.html)

# 부록

패키지 및 서버명	버전
Windows	1.1
Python	3.11.5
Selenium	4.16.0
Webdriver-manager	4.0.1
pandas	2.0.3
numpy	1.24.3
seaborn	0.12.2
matplotlib	3.7.2
sklearn	1.11.1
scipy	1.3.0

[시스템 요구사항 확인 내용]



프로젝트 진행 기간

2024.01.15 ~ 2024.01.19 15:00 (총 5일)

1조 팀원

심현지, 이신영, 이인철, 이효준, 조용재

# 온라인 IT교육 유입 요인 분석