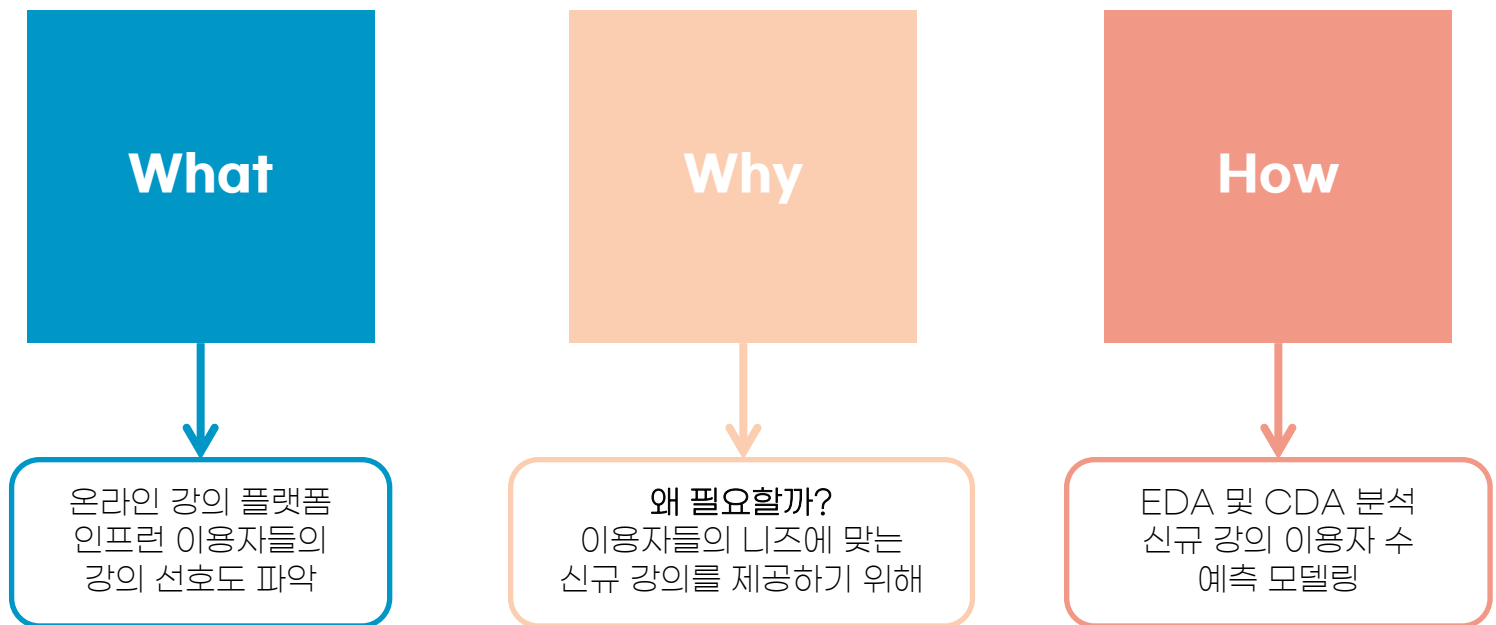




**사람들은  
어떤 강의를  
듣고 싶을까?**

# 문제 정의

---



A 이용자가 데이터 분석가가 되기 위해 데이터 수집에 관한 기초 강의 콘텐츠를 들었다고 가정한다.

수강 후, 실제 적용이 어려워서 중간 수준의 데이터 수집 강의를 듣고 싶을 수 있다.

A가 원하는 강의를 제공하지 못하는 경우, 이탈할 수 있으며 이는 궁극적으로 매출에 영향을 미치게 된다.

따라서 이용자들이 원하는 강의를 파악하고 제공하기 위해 분석을 실시하였다.

# 데이터 수집

## 수집 대상

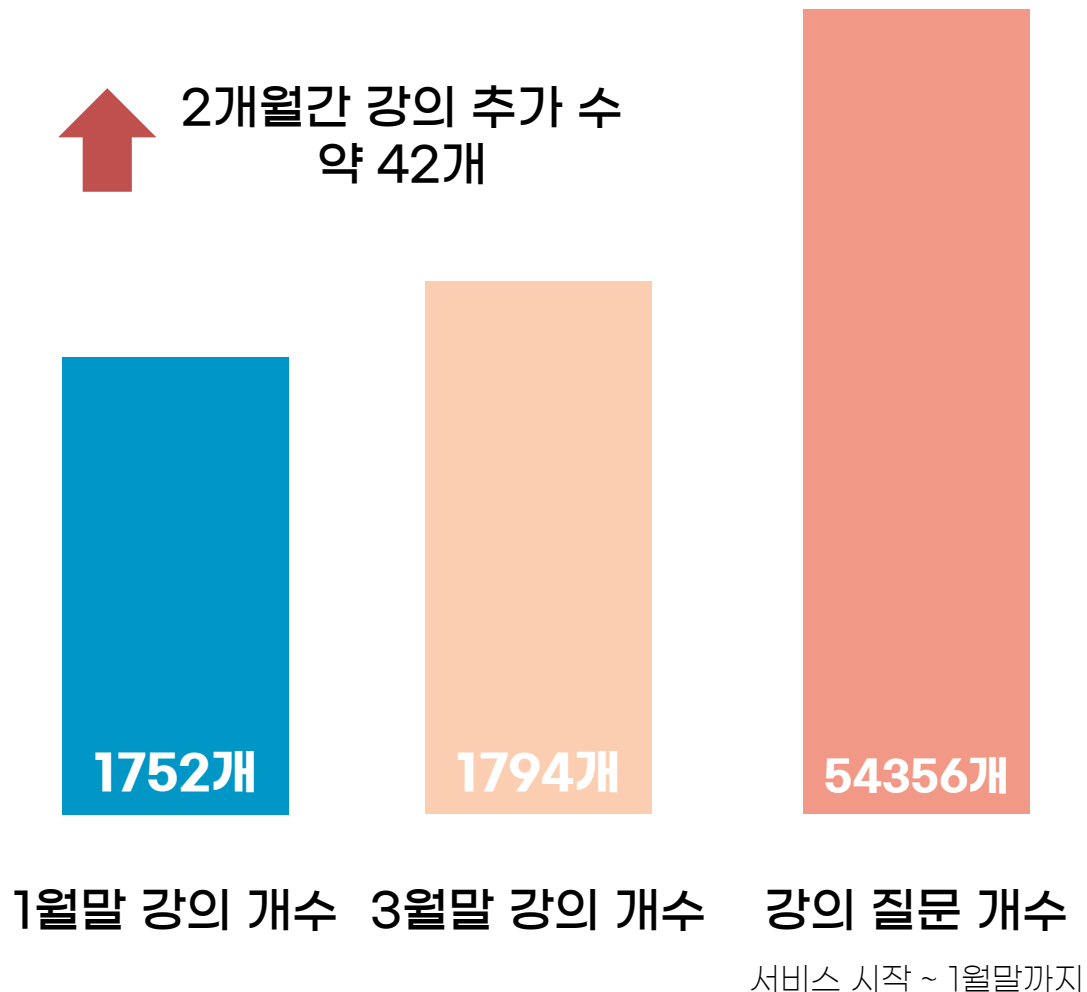
강의 / 질문 게시판

## 수집 방법

셀레니움 사용

## 질문 게시판 수집 이유

이용자가 질문을 작성하였을 때,  
강사의 답변율이 높다면  
해당 강의를 이용자들은 더 선호하는지를  
파악하기 위해 질문 데이터를 수집하였다.



# Data Description

## 강의 게시판

### 강의 분야

1. 메인 카테고리
2. 서브 카테고리

### 강의 정보

3. 강의 이름
4. 강사 이름
5. 강의 관련 태그
6. 강의 섹션 수
7. 수료증 발급 유무
8. 강의 난이도

### 강의 가격

9. 일반 가격
10. 현재 할인율
11. 현재 할인 가격

### 강의 시간

12. 총 강의 시간
13. 수강 가능 기간

### 강의 평가

14. 강의 리뷰 점수
15. 강의 리뷰 개수
16. 수강생 수

## 질문 게시판

### 질문 정보

1. 질문 게시 글 제목
2. 질문자
3. 질문 날짜
4. 질문 내용
5. 질문한 강의 내용
6. 질문한 강의 섹션

### 답변 정보

7. 답변자 목록
8. 답변 날짜 목록
9. 답변 내용 목록

# Feature Engineering

## 강의 게시판

### 강의 분야

1. 메인 카테고리
2. 서브 카테고리

### 강의 정보

3. 강의 이름
4. 강사 이름
5. 강의 관련 태그
6. 강의 섹션 수
7. 수료증 발급 유무
8. 강의 난이도

### 강의 가격

9. 일반 가격
10. 현재 할인율
11. 현재 할인 가격

### 강의 시간

12. 총 강의 시간
13. 수강 가능 기간

### 강의 평가

14. 강의 리뷰 점수
15. 강의 리뷰 개수
16. 수강생 수

### 강의 질문

17. 총 질문 수
18. 강사의 총 답변 수
19. 강사의 답변율
20. 답변까지 걸린  
기간 평균

## 질문 게시판

### 질문 정보

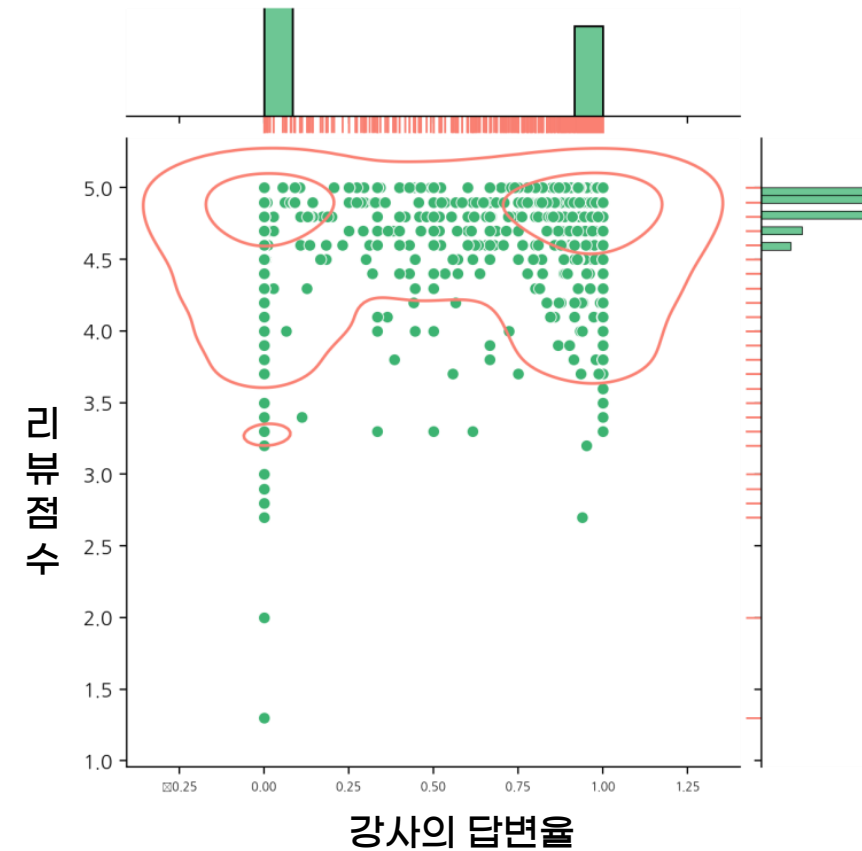
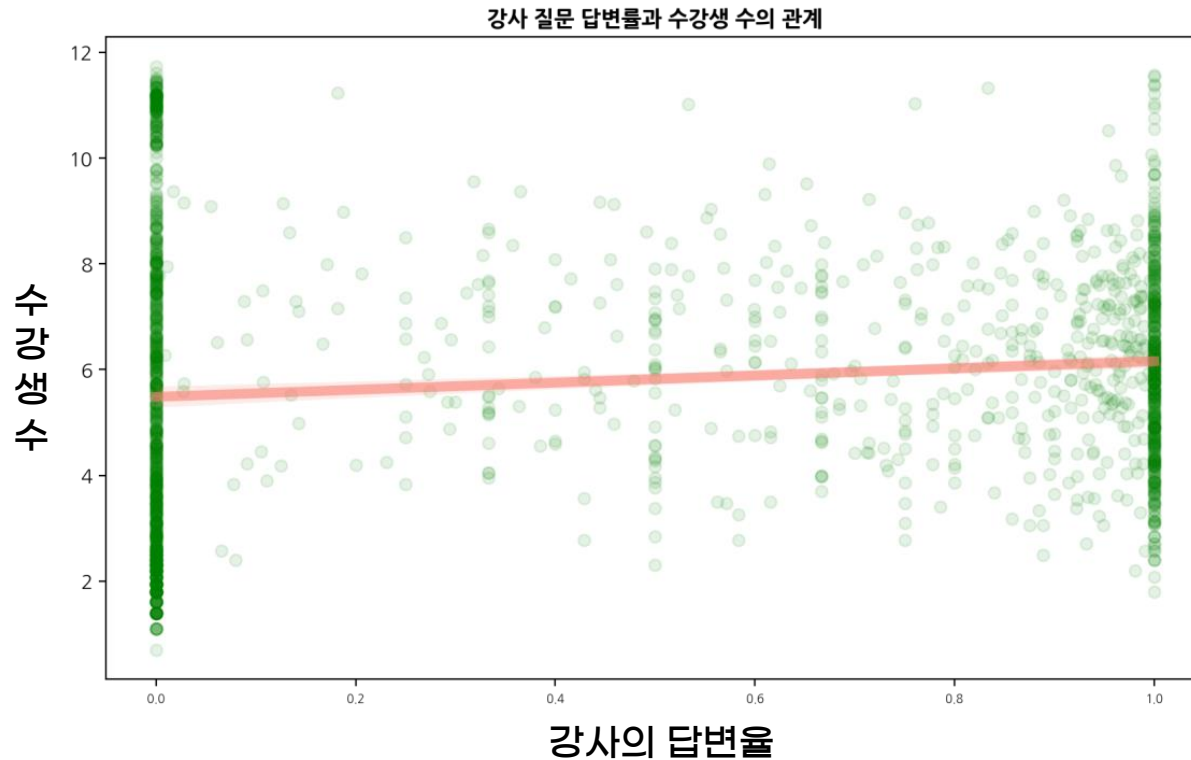
1. 질문 게시 글 제목
2. 질문자
3. 질문 날짜
4. 질문 내용
5. 질문한 강의 내용
6. 질문한 강의 섹션

### 답변 정보

7. 답변자 목록
8. 답변 날짜 목록
9. 답변 내용 목록

**질문 게시판 데이터를 정제하여  
강의 게시판 데이터에 변수 추가**

# 강의와 질의답변의 관계



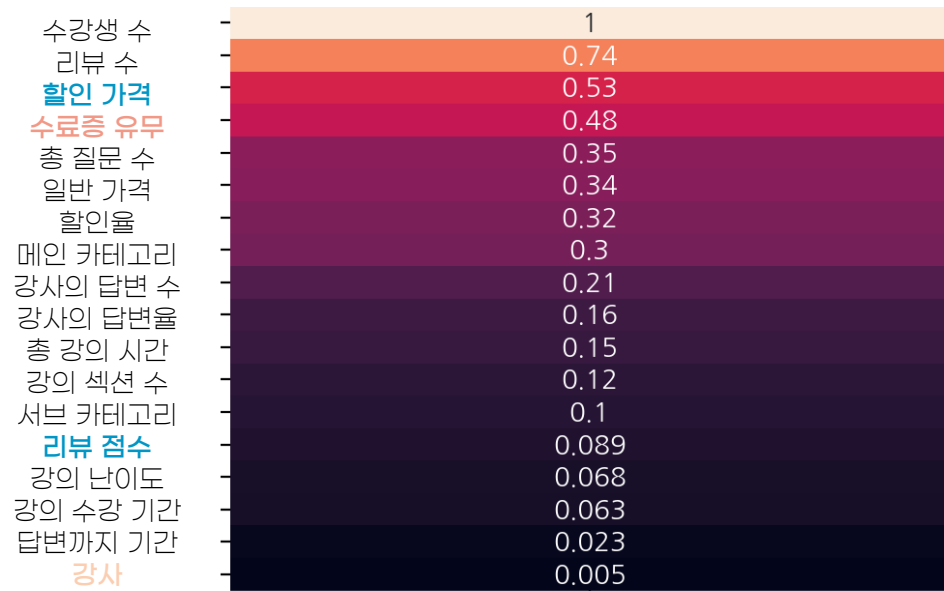
수강생 수와 강사의 답변율의 관계 : 강의별 강사의 답변율은 아예 하지 않거나 모든 질문에 답변한 것으로 나타난다.  
-> 강사의 답변 유무에 따라 강의를 더 선호하지는 않는 것으로 관찰되었다.

리뷰 점수와 강사의 답변율의 관계 : 강의별 강사의 답변율은 리뷰 점수와의 영향이 적었다.

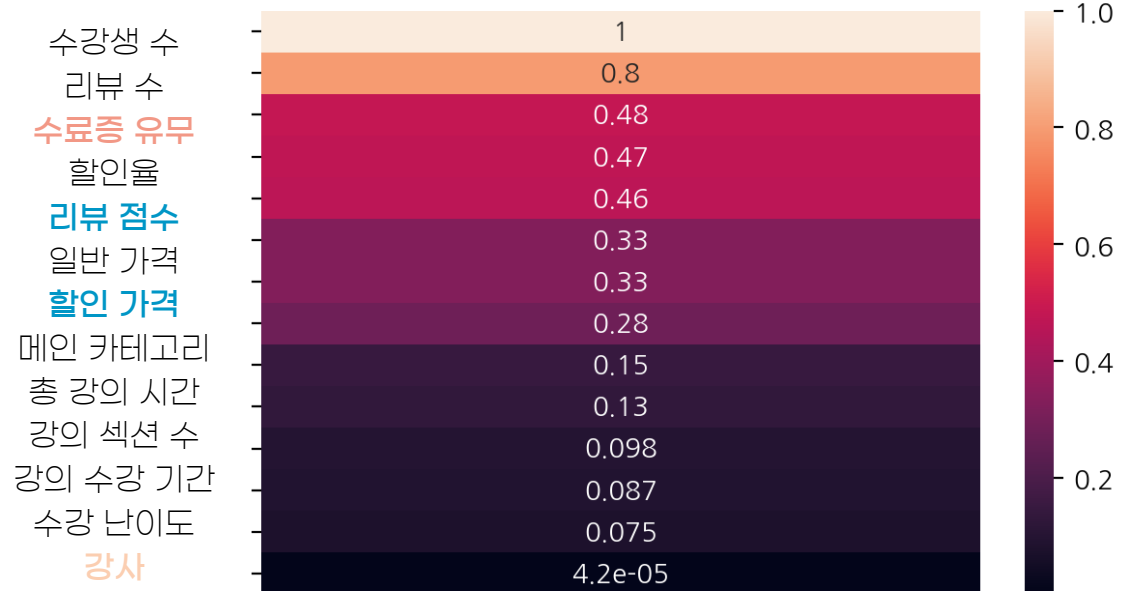
데이터 수집 단계에서 세운 가설인 **질의답변에 따라 강의 선호도가 높을까?** 는 기각해야 함을 관찰하였다.  
- 따라서 3월말 데이터 수집에서는 질문 게시판 데이터를 수집을 진행하지 않았다.

# EDA - 상관관계 분석

강의 선호도를 파악하기 위해 강의별 수강생 수를 종속변수로 지정하여 EDA를 진행한다.



1월말 기준 상관관계



3월말 기준 상관관계

**할인 가격**과 **리뷰점수** : 시간이 지남에 따라 할인 가격 지표는 줄어들고 리뷰 점수의 지표는 증가하였다.

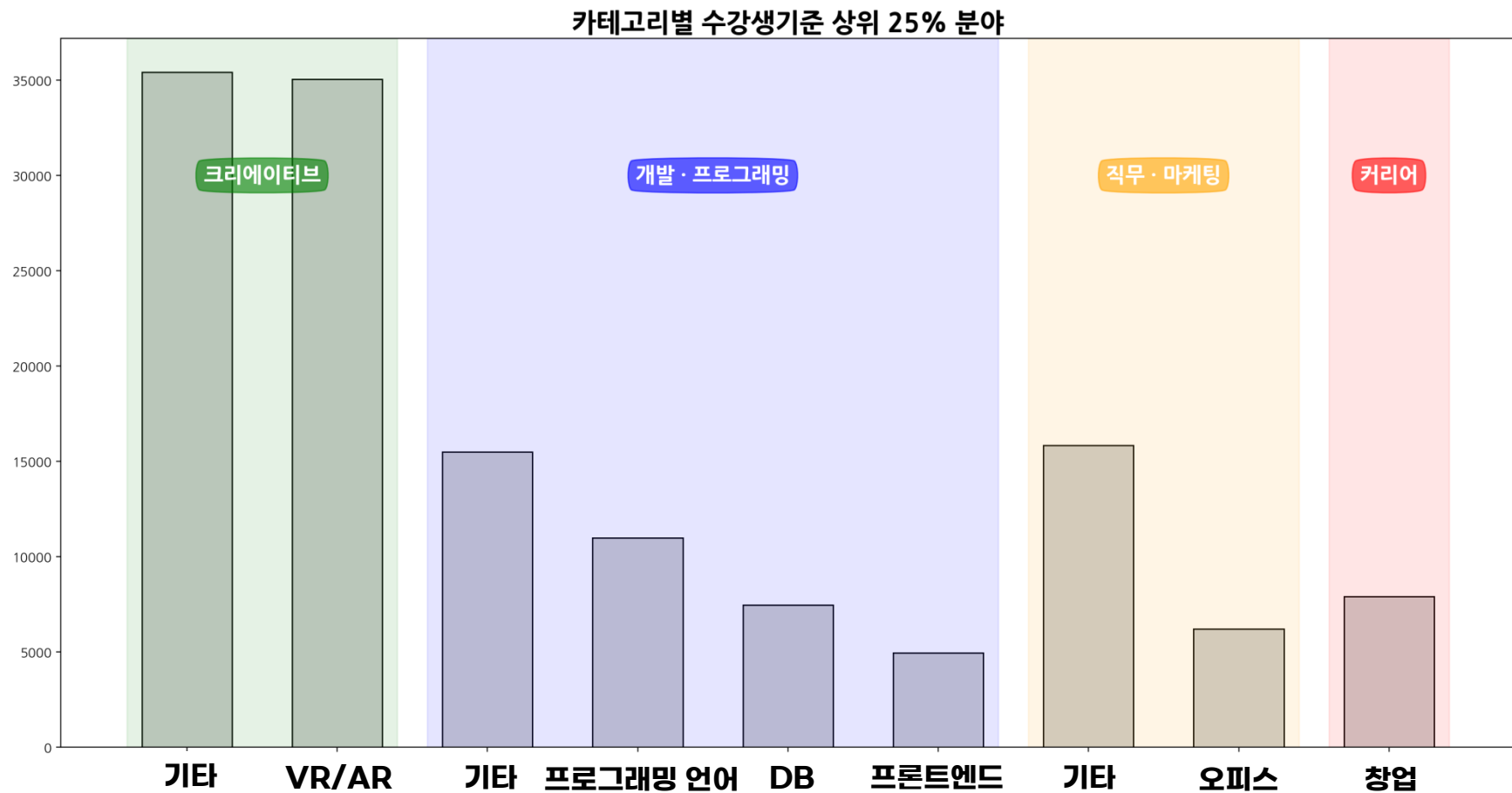
-> 1월에는 많은 강의를 0원에 제공하는 마케팅을 진행하였다. 마케팅으로 인해 할인된 강의를 선호한 것으로 보이나, 마케팅 이후에는 할인 가격보다 리뷰 점수와 상관관계가 증가한 것으로 보인다.

**수료증 유무** : 수강생 수와 밀접한 관계를 보임 -> **가설 1. 수료증 발급은 수강에 큰 영향을 준다?**

**강사** : 도메인 지식을 통해 온라인 강의 플랫폼은 유능한 강사에 수강생이 집중된다고 알고 있었으나 상관관계에서는 미미한 관계를 보였다.

-> **가설 2. 수강생이 강의를 선택할 때 강사는 중요하지 않다?**

# EDA - 시각화 분석 (카테고리)

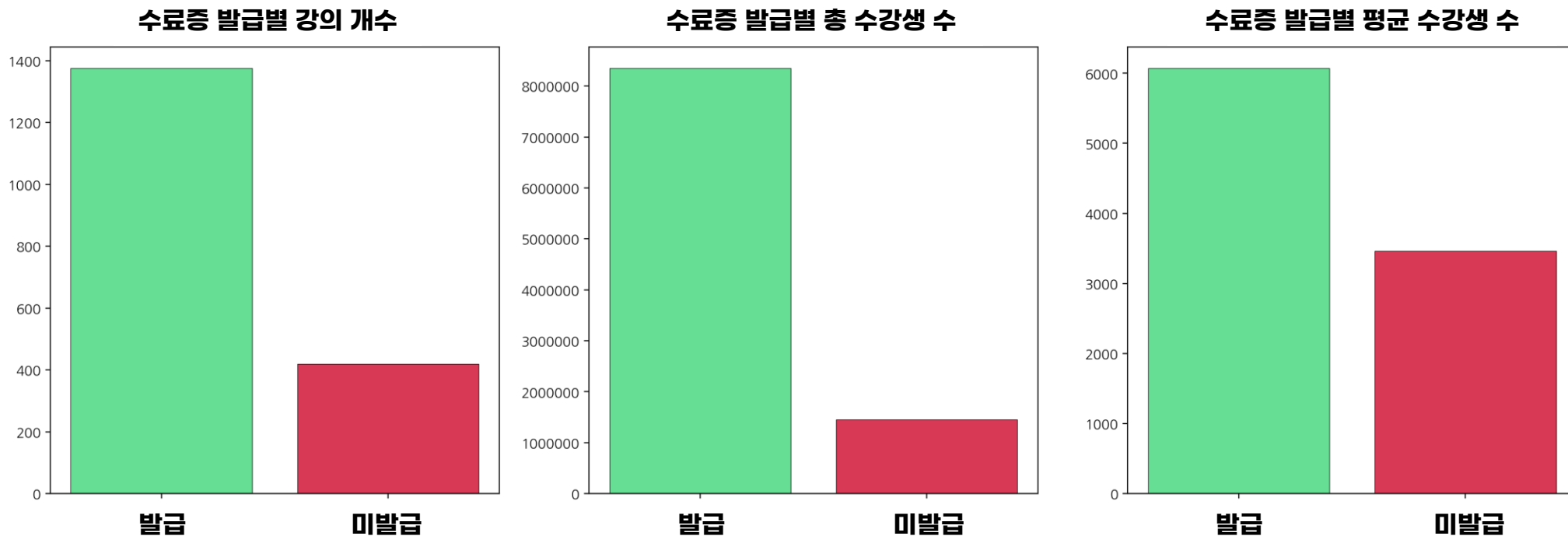


상단에 표기된 카테고리는 메인 카테고리를 의미한다. 하위 서브 카테고리 중, 메인 카테고리에서 수강생 수를 기준으로 상위 25%의 서브 카테고리를 표기하였다.

각 메인 카테고리에서 **기타** 카테고리가 가장 높은 수강생 수를 보였다. -> 기타 카테고리 중에 인기 있고 공통적인 카테고리로 묶을 수 있다면 신규 카테고리를 개설하여 이용자들이 쉽게 접근할 수 있도록 개선이 필요해 보인다.



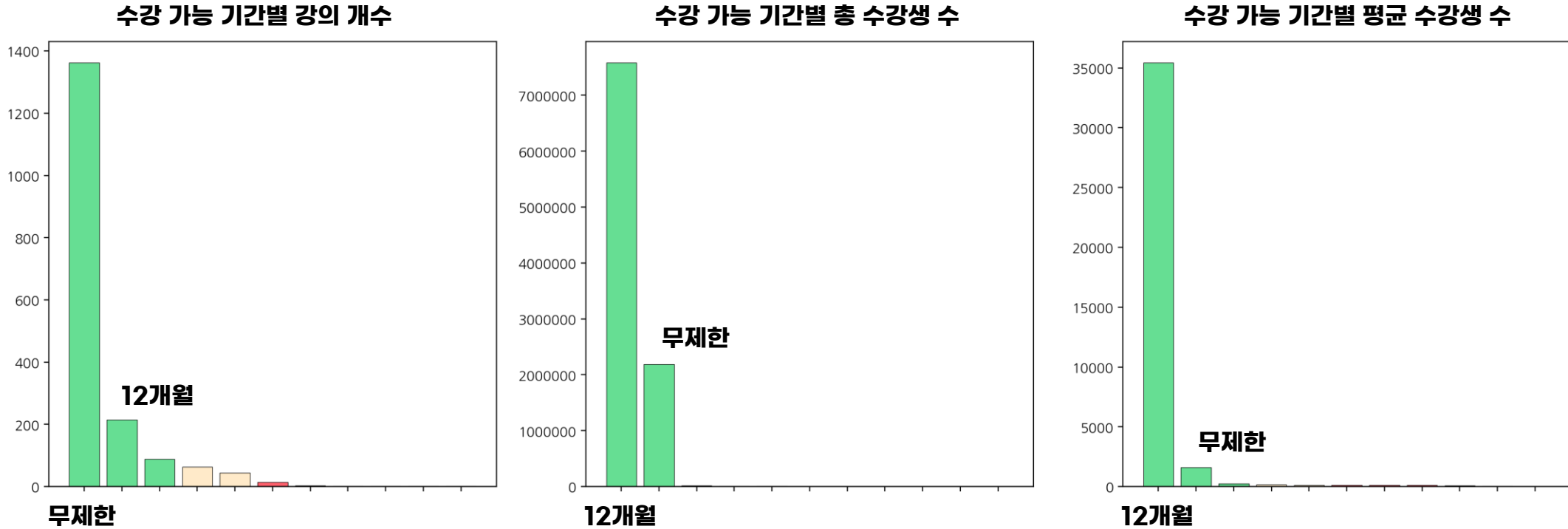
# EDA - 시각화 분석 (수료증 발급 유무)



수료증을 발급하는 강의의 수와 평균 수강생 수 모두 미발급 강의에 비해 높았다.

상관관계 분석에서 수료증 유무 변수가 수강생 수와 밀접한 관계를 보이는 것은 알았지만, 시각화 분석을 통해 수료증을 **발급**하는 강의의 수강생 수가 높은 것을 관찰하였다.

# EDA - 시각화 분석 (수강 가능 기간)

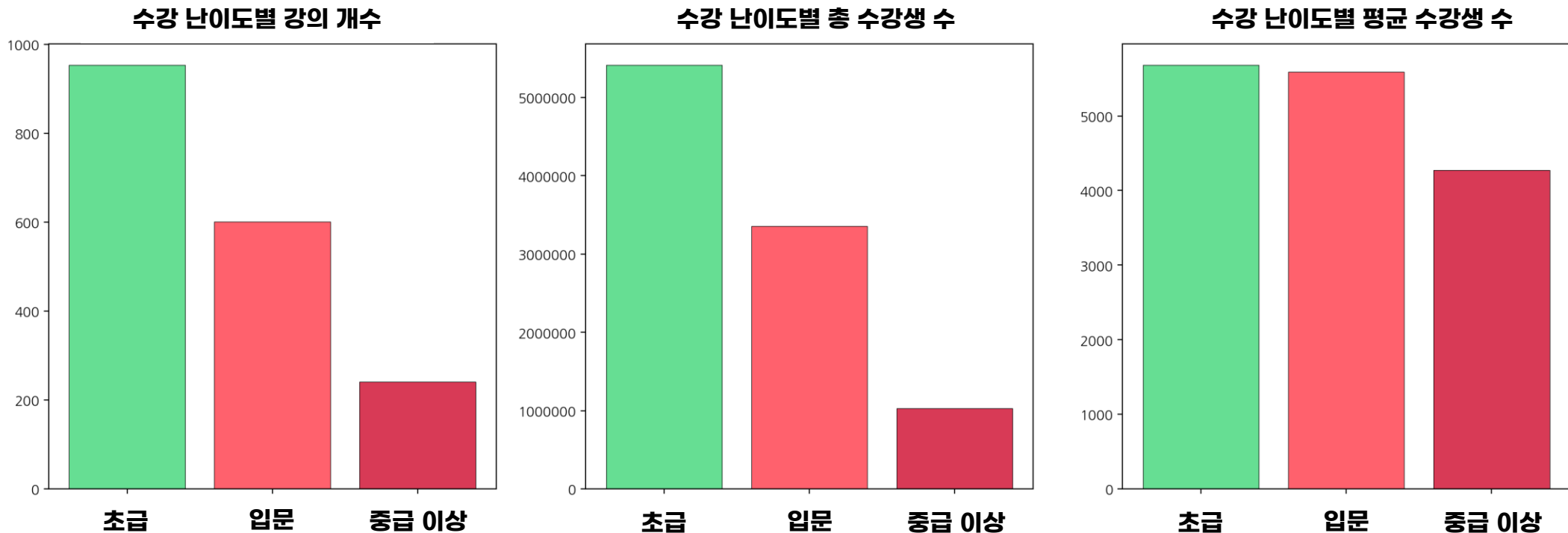


수강 가능 기간은 수강 시작 날짜를 기준으로 강의 서비스를 이용할 수 있는 기간을 말한다.

무제한으로 제공되는 강의의 수가 많았지만 총 수강생수와 평균 수강생 수는 12개월 제한인 강의를 매우 높았다.

-> **가설 3. 무제한으로 제공되는 강의보다 수강 기간 제한이 있는 강의들이 더 높은 퀄리티를 제공한다?**

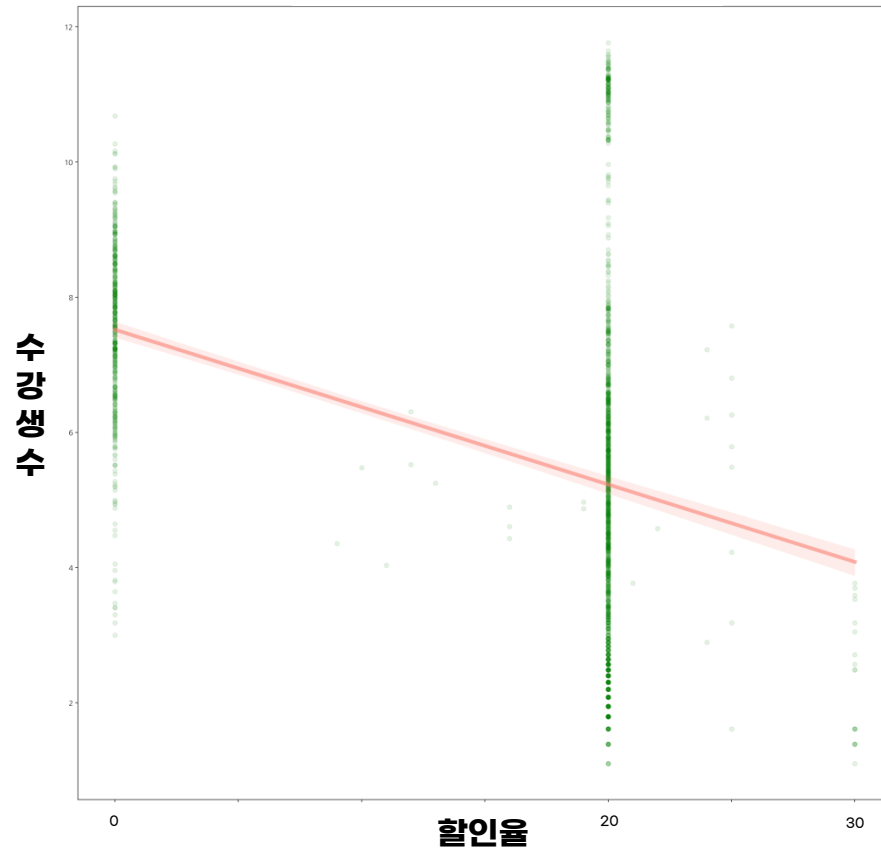
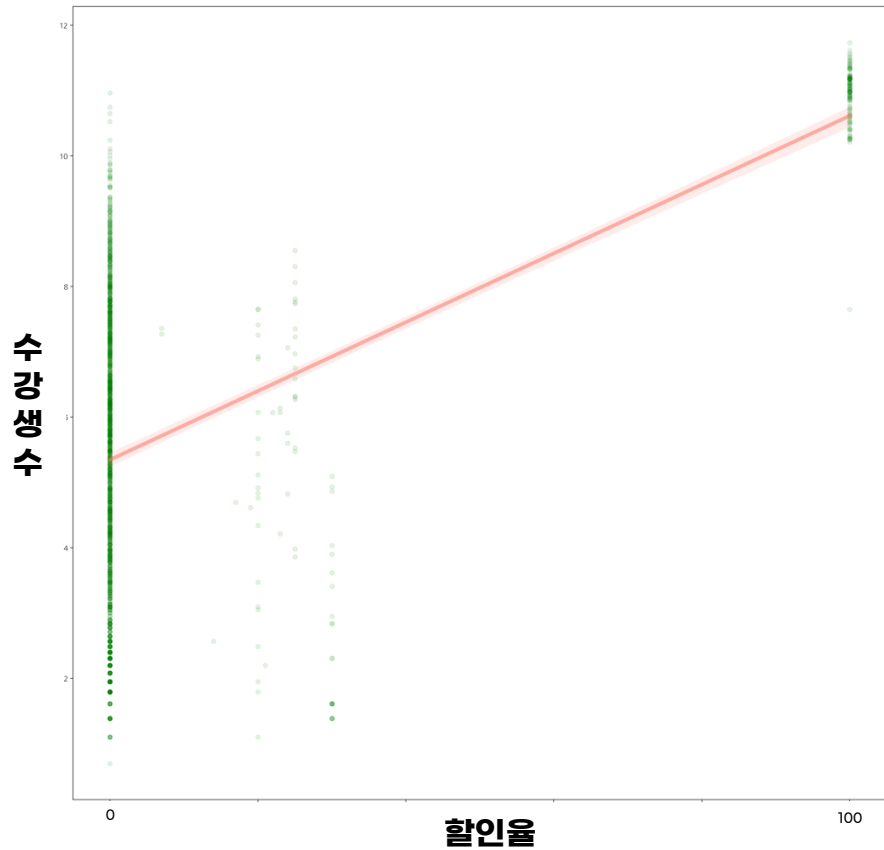
# EDA - 시각화 분석 (수강 난이도)



수강 난이도는 초급의 강의 수, 총 수강생 수가 많았다. 인프런을 이용하는 고객들은 초급 난이도의 강의를 선호하는 것으로 관찰되었다.

평균 수강생 수를 보았을 때, 난이도별 강의 개수 및 수강생 수가 고루 분포되어 있는 것으로 보아, 공급과 수요의 균형이 잘 이루어져 있었다.

# EDA - 시각화 분석 (할인율)



1월에는 할인율 100%로 여러 강의들을 제공하는 마케팅을 진행하여 이용자들의 많은 사랑을 받았다.

3월에는 1월에 무료로 제공된 강의와 수강생 수가 적은 강의들에 대해 20%의 할인을 제공하고 있는 것으로 관찰된다.

# EDA - 키워드 분석 (강의 제목 키워드)

Top5 강의 제목 키워드	1월	3월	증가수
파이썬	157	155	-2
데이터	102	108	6
웹	94	99	5
게임	83	82	-1
앱	73	73	0

강의 제목 키워드 수

Top5 강의 제목 키워드	1월	3월	증가수
파이썬	907,880	956,013	48,133
데이터	216,603	238,849	22,246
웹	147,577	161,335	13,758
게임	140,405	148,834	8,429
앱	68,134	71,886	3,752

강의 제목 키워드별 수강생 수

1월과 3월을 비교하였을 때, 회사는 **데이터**와 **웹**에 대한 신규 강의를 중점적으로 추가하였으며, 이용자들은 **파이썬**, **데이터**, **웹** 순으로 수강생이 증가하였다.

강의 제목 키워드 수는 **회사**의 입장에서 이용자들이 선호할 것이라고 생각하는 강의 목록을 알 수 있고, 반대로 강의 제목 키워드별 수강생 수는 **이용자**들이 선호하는 강의임을 알 수 있다.

# EDA - 키워드 분석 (강의 태그 키워드)

Top5 강의 태그 키워드	1월	3월	증가수
웹	198	208	10
데이터	192	197	5
파이썬	136	134	-2
자바스크립트	91	98	7
MS-오피스	89	95	6

강의 태그 키워드 수

Top5 강의 태그 키워드	1월	3월	증가수
MS-오피스	379,598	394,092	14,494
자바스크립트	315,463	338,735	23,272
웹	248,458	265,663	17,205
파이썬	153,354	164,055	10,701
데이터	86,292	93,628	7,336

강의 태그 키워드별 수강생 수

1월과 3월을 비교하였을 때, 회사는 웹, 자바스크립트, MS-오피스 순으로 신규 강의를 중점적으로 추가하였으며, 이용자들은 MS-오피스, 자바스크립트, 웹 순으로 수강생이 증가하였다.

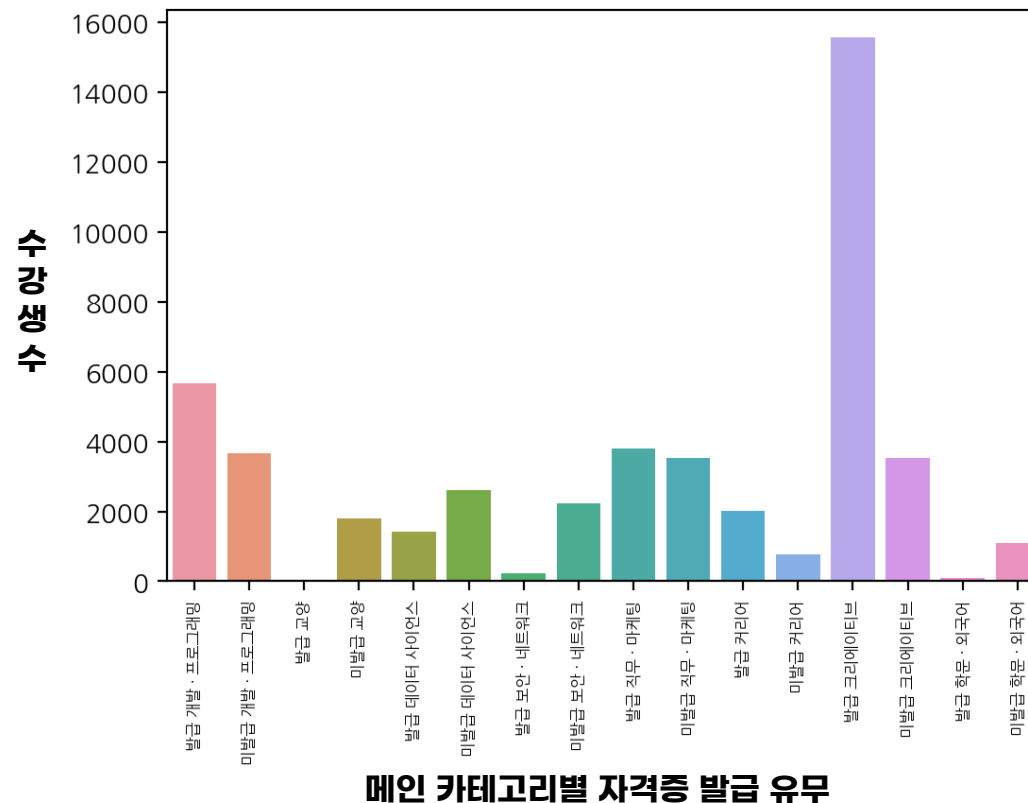
강의 제목과는 다르게 강의 태그 기준 키워드는 회사와 이용자의 선호도가 달랐다.  
제공되는 MS-오피스와 자바스크립트의 강의는 상대적으로 적으나, 웹, 파이썬, 데이터에 비해 많은 수강생을 보유하고 있다.

MS-오피스와 자바스크립트의 수요가 많은 만큼, 해당 태그와 관련된 신규 강의를 개설하는 것을 고려해야 할 것으로 보인다.

# CDA - 가설 1. 수료증 발급 유무는 수강에 영향을 미친다?

## 가설1 검증 프로세스 : 통계치 분석 -> 시각화 분석

수료증 발급 유무	미발급 평균	발급 평균
수강생 수	3,302명	5,943명
평균 일반 가격	730원	56,391원
평균 할인 가격	730원	51,349원



수료증 발급 유무에 따라 변수들의 통계치를 분석하였다.

수료증을 발급하는 강의가 수강생 수도 많았고 가격도 비쌌다. 가격이 비쌌음에도 수료증을 발급하는 강의를 선호하는 이유를 시각화 분석을 통해 알아보았다.

수료증 발급을 기준으로 메인 카테고리별 수강생 수를 시각화한 결과, **크리에이티브** 분야에서 수료증 발급에 대한 선호도가 굉장히 높았다.

즉, 가격이 더 비쌌음에도 특정 분야들에서 유의미한 수강생 수의 차이를 보였으므로 **가설 1은 실제 상관관계를 갖는 것을 관찰**하였다.

# CDA - 가설 2. 수강생이 강의를 선택할 때 강사는 중요하지 않다?

## 가설2 검증 프로세스 : 변수 인코딩 -> 모델 선택 -> 변수 중요도 분석

변수 이름	변수 중요도	변수 이름	변수 중요도	변수 이름	변수 중요도
리뷰 수	0.75580	리뷰 수	0.65460	리뷰 수	0.64504
강사_유용한T학습	0.58681	강사_유용한T학습	0.48447	강사_유용한T학습	0.49658
할인 가격	0.06284	일반 가격	0.05701	일반 가격	0.13950
일반 가격	0.05547	할인 가격	0.03743	총 강의 시간	0.01786
메인 카테고리_크리에이티브	0.03244	메인 카테고리_크리에이티브	0.01536	메인 카테고리_크리에이티브	0.01358
총 강의 시간	0.02392	총 강의 시간	0.01451	강사_인프런	0.01026
강의 가능 기간	0.01617	강사_인프런	0.01064	할인율	0.00947

Random Forest 변수 중요도

GB 변수 중요도

XGB 변수 중요도

상관관계 분석에서는 강사와의 상관관계가 없다고 분석되었으나 실제 수강생 수 예측 모델을 생성한 결과, **강사** 변수의 중요도가 굉장히 높았다.

특히, **유용한T학습**이라는 강사는 리뷰 수와 함께 모델에서 가장 영향력 있게 데이터를 분류하고 예측할 수 있는 변수로 꼽히고 있다.

상관관계 분석과 변수 중요도 분석을 종합하여 볼 때, **가설 2는 인기 있는 몇몇의 강사들은 이용자들이 강의를 선택할 때 중요한 변수로 작용**하고 있다.



# CDA - 가설 3. 수강 기간 제한이 있는 강의가 더 높은 퀄리티를 가질까?

---

## 가설3 검정 프로세스 : 통계치 분석

---

통상적으로 이용자 입장에서는 무제한으로 제공되는 강의를 선호할 것으로 생각되지만, 시각화 분석에서는 12개월로 제한된 강의의 수강생 수가 가장 많았다. 그 이유를 통계치 분석을 통해 관찰하였다.

수강 가능 기간별 통계치를 살펴본 결과, 수강 가능 기간은 **할인율**에서 큰 차이를 보였다.

이는 1월에 마케팅 이벤트를 통해 **무료로 제공한 강의**들이 12개월의 수강 가능 기간을 갖고 있었기 때문으로 관찰되었다.

수강 기간 제한에 따라 강의를 선택하는 기준이 달라지는 것이 아니라 할인한 강의에 대한 수강생 수 증가이므로 **가설 3은 바람직하지 않은 것으로 관찰**된다.

# 모델링 : 신규 강의의 수요(수강생 수 기준) 예측 모델 생성

---

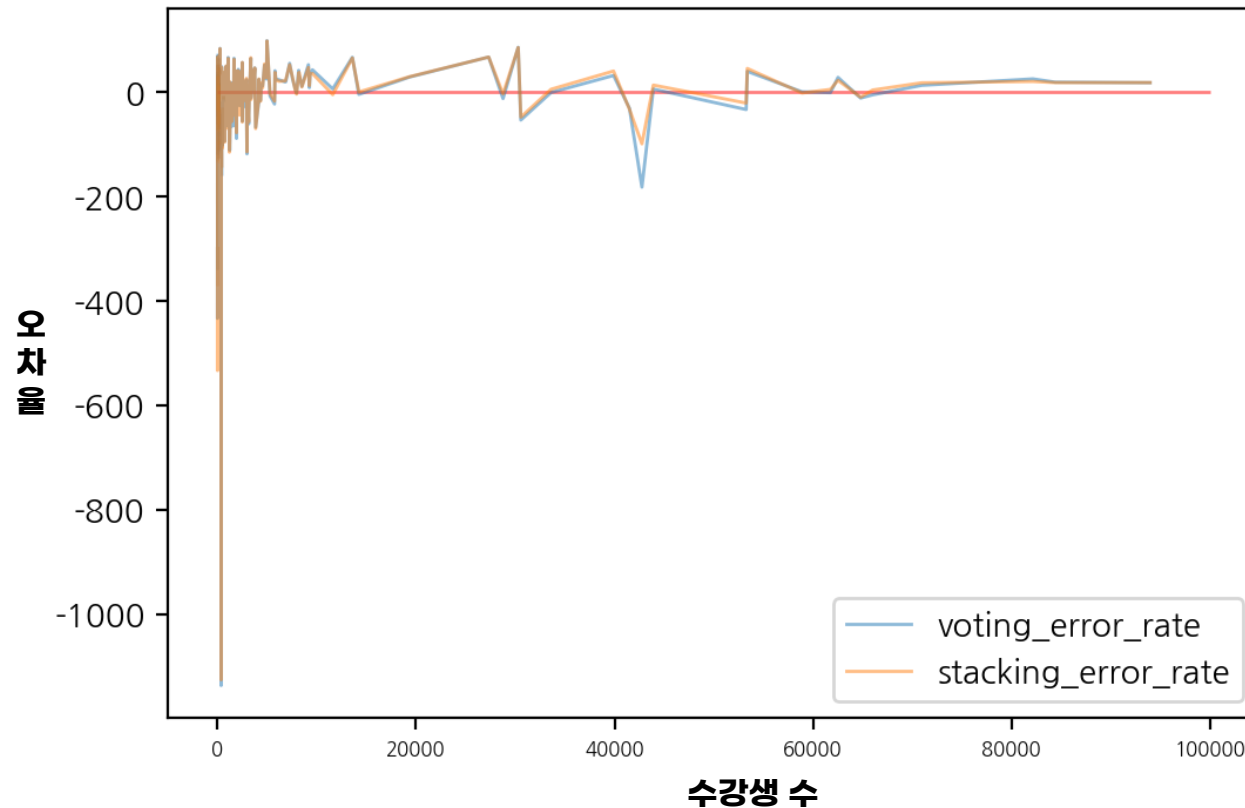
## 모델링 설정

사용 모델	Ridge, Lasso, Elastic Net, Random Forest, Gradient Boosting, Extra Gradient Boosting
평가 지표	RMSE
교차 검증	5회
하이퍼 파라미터 튜닝	GridSearchCV
학습, 테스트 데이터 분리	<code>train_test_split(Train_size = 0.8, Test_size = 0.2, Shuffle = True)</code>
앙상블 기법	Voting, Stacking

# 모델링 결과

앙상블 기법	Train Data	Test Data
Voting	0.2586	0.4941
Stacking	0.2892	0.4830

\* 평가지표 : RMSE



두 앙상블 모두, 약간의 overfitting이 있었지만 앙상블 전보다 수강생 수(종속 변인)을 더 잘 예측했다.

우측 그림처럼 수강생 수가 적은 강의들은 오차율이 매우 컸으며, 수강생 수가 많은 강의들은 비교적 잘 예측하였다.

변수 중요도 분석에서 살펴보았듯이 모델들은 **리뷰 수**와 **강사\_유용한T학습**에 의해 대부분의 데이터들을 분류하므로 해당 데이터만으로 신규 강의의 수요(수강생 수) 예측에는 한계점이 나타났다.

# 인사이트

---

- 메인 분야별 **기타** 카테고리의 세분화 및 신규 강의 개설 필요
- 무료로 강의를 제공하는 이벤트에 참여한 강의들이 그렇지 않은 강의들에 비해 **약 44배** 많은 수강생 수를 가짐
  - > **신규 이용자 유입에 효과적**
- 강사의 질의답변은 강의 수강에 큰 영향을 미치지 않음
- 강의 제목에 **파이썬 / 데이터 / 앱**이라는 단어가 포함된 강의를 선호하였음
- 강의 관련 태그에 **MS-오피스 / 자바스크립트 / 웹**이 포함된 강의를 선호하였음
- **크리에이티브** 및 **개발 프로그래밍** 카테고리를 이용하는 고객들은 수료증 발급하는 강의를 선호함
- 여러 강의 서비스를 제공하는 강사의 경우, 해당 강사의 강의별 수강생 수는 유사하였음
- 유용한IT학습이라는 강사가 전체 수강생 수에 **약 78%**를 차지하고 있음
  - > 온라인 강의 플랫폼 특성상 유명하거나 유능한 강사가 많은 지분을 차지하고 있지만,  
너무 한 명의 강사에 집중된 현상은 위험 요소를 갖고 있다. 따라서 다른 **강사들의 추가 영입이 필요**해 보임

# 데이터 분석의 한계점

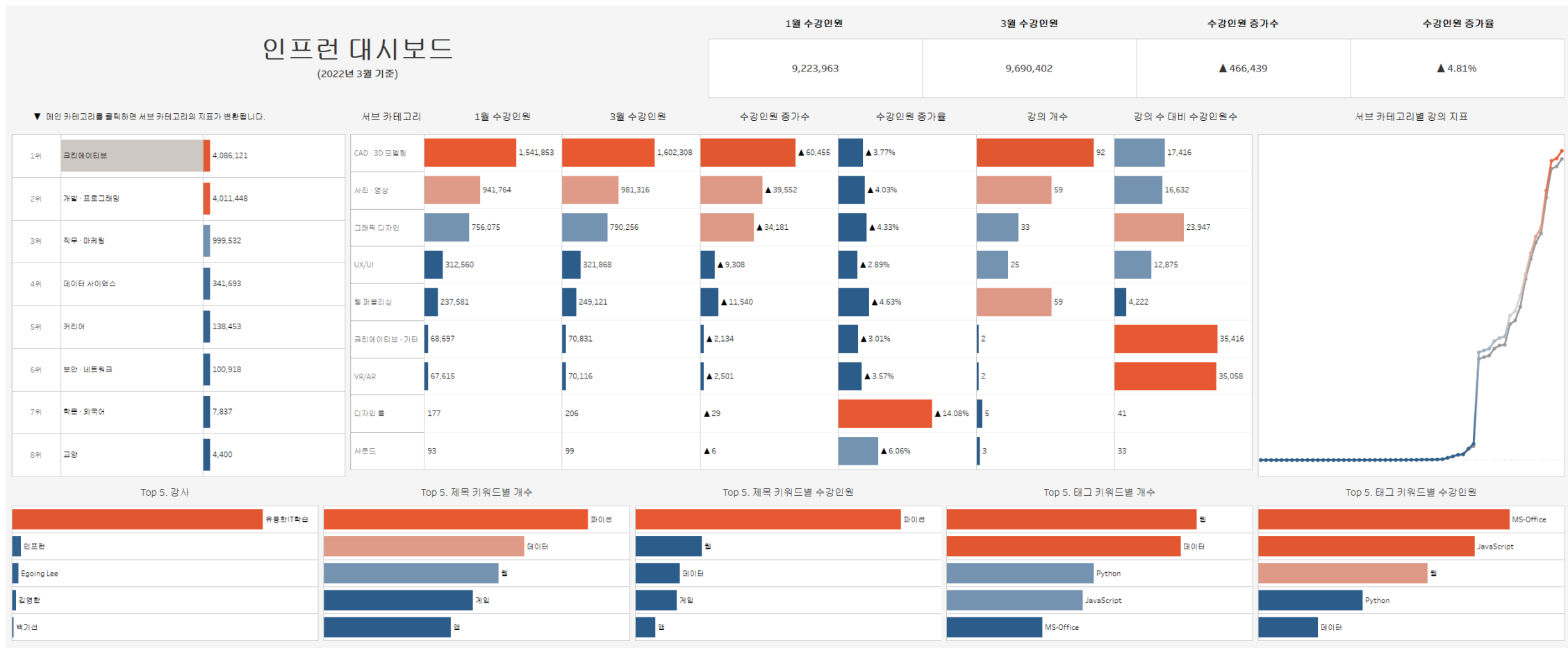
- 서비스 플랫폼에서 스크래핑한 데이터만으로는 신규 강의의 수요 예측이 어려움
- 수집한 데이터는 강의를 기준으로 수집되었으나, **이용자들의 데이터**를 수집할 수 있다면, 고객 세분화 분석, 이탈 분석, 예측 모델링 등 다양한 분석이 가능해짐

## 더 나은 분석을 위해 필요해 보이는 데이터


\* 이용자 기준

1. 나이
2. 성별
3. 전공
4. 서비스 이용 시간
5. 서비스 접속 간격
6. 검색 키워드
7. 서비스 구매 목록
8. 학습중인 강의
9. 학습 완료한 강의
10. 수료증 목록
11. 질문 유무
- ⋮

# 대시보드 제작



- 상단 : 전체 이용자 성장율
- 중단 : 메인 & 서브 카테고리별 성장율
- 하단 : 강사별 수강생 수 / 제목 & 태그 키워드별 수요 및 공급량



**Thank you  
for viewing  
my portfolio**