

데이터 분석 프로젝트

패스트캠퍼스 결제 데이터 분석

강준석

프로젝트 개요

- **목표**

- 패스트캠퍼스 2022년 온라인 강의 결제내역을 통해 인사이트 도출

- **프로젝트 기간**

- 23.05.09. ~ 23.05.16. (일주일)

- **데이터**

- 결제 데이터 15,328 개 (1인 중복결제, 프로모션 결제 포함)

탐색적 데이터 분석(EDA)

• 데이터 분석


```
RangeIndex: 159328 entries, 0 to 159327
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   거래id      159328 non-null  int64
 1   유형        159328 non-null  object
 2   고객id      159328 non-null  int64
 3   코스ID      159328 non-null  int64
 4   사이트      159328 non-null  object
 5   포맷        159328 non-null  object
 6   카테고리    159328 non-null  object
 7   코스(상품) 이름  159328 non-null  object
 8   거래일자    159328 non-null  object
 9   쿠폰이름     159328 non-null  object
10   판매가격     159328 non-null  object
11   결제수단     159328 non-null  object
12   실거래금액   159328 non-null  int64
13   쿠폰할인액    159328 non-null  object
14   거래금액     159328 non-null  object
15   환불금액     159328 non-null  object
dtypes: int64(4), object(12)
memory usage: 19.4+ MB
```

- 활용할만한 데이터 선별
 - 거래일자 : 시계열 데이터 (결제 시점)
 - 고객 Id : 결제 주체 특정 가능
 - 코스(상품) 이름
 - 카테고리
 - 이외 Column 제거 이유
 - 판매금액, 쿠폰이름 등
 - 프로젝트 특성상 데이터에 대한 세부소통이 불가능한 상황에서 애매한 value의 컬럼 제외
- e.g. 쿠폰의 경우 쿠폰마다 의도가 다르고 목적이 모호한 value 존재

탐색적 데이터 분석(EDA)

• 회사 분석

산업과 교육 간 미스매치를 해소하는 교육 콘텐츠 회사 '패스트캠퍼스...
인생과 커리어를 바꾸기 위해 가장 먼저 찾는 선택지, 패스트캠퍼스입니다.

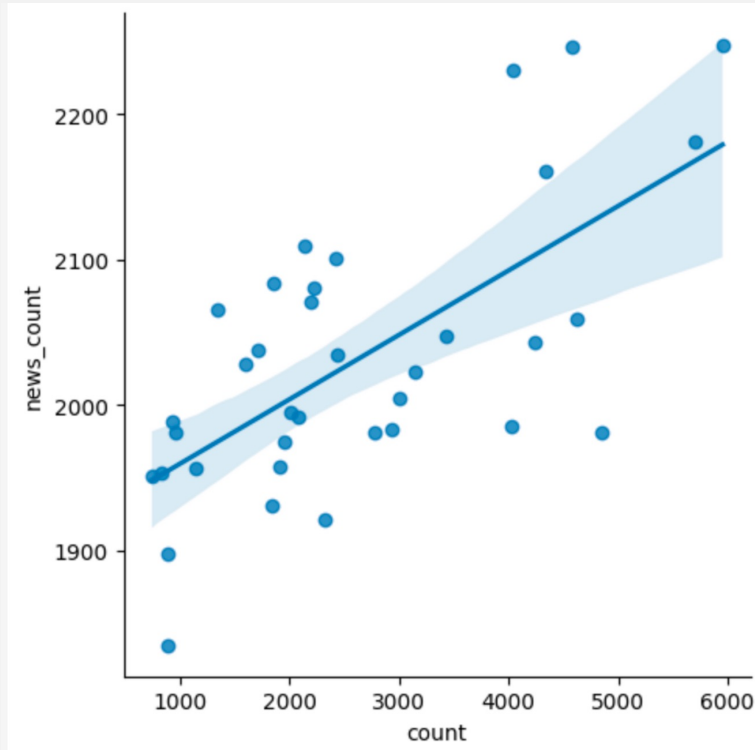
 <https://youtu.be/sENMbS6dvA0>



- Life-Changing Education : 미래를 바꾸는 교육 콘텐츠 회사 지향
- 산업과 교육의 괴리를 없애주는 최신성
- 타 기사에서도 “타임 투 마켓” 및 “콘텐츠 라이프 사이클” 등을 강조
 - 두 가지 모두 학습 수요의 타이밍, 이미 제작한 콘텐츠도 2년에 한번 새로 제작하는 등 최신 흐름을 담고, 유행을 아는 것을 중요하게 생각함.

탐색적 데이터 분석(EDA)

• 최신성 관련 상관관계 분석



- 강의 카테고리별 결제빈도와 근 6개월 간(결제 기준) 카테고리명 뉴스 상관관계 시각화
- 상관관계가 명확히 양의 상관관계를 가졌으나 데이터를 통한 향후 액션이 떠오르지 않아 인사이트가 아니라고 판단
- 해당 데이터를 보고 떠오르는 액션이 없다면 현업에서 필요한 데이터가 아니며, 몇가지 기준을 새로 세우고 프로젝트를 임하게 됨.

프로젝트 기준 확립

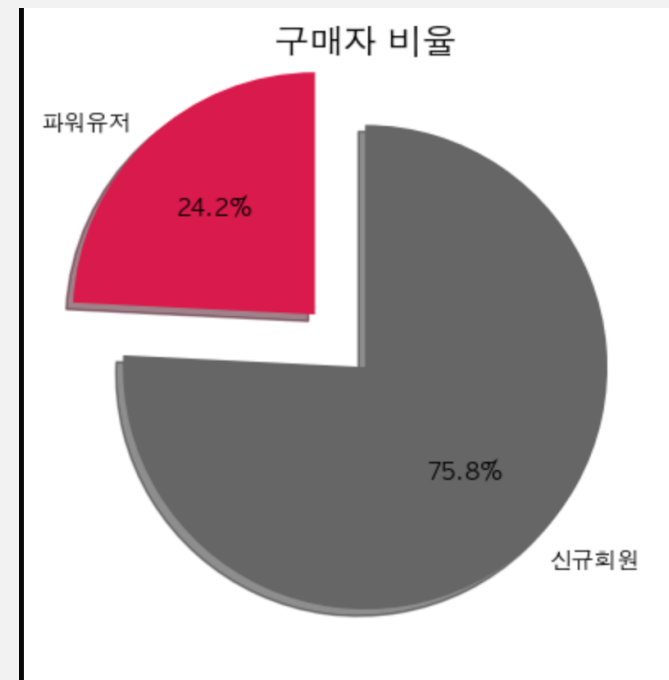
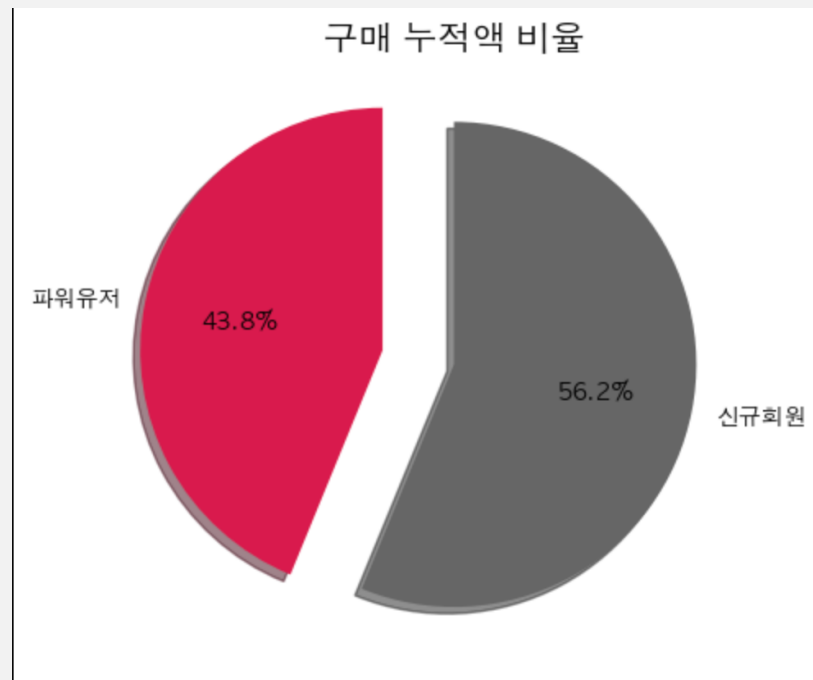
- 원칙

1. 인사이트 자체가 프로젝트 데이터에서 완결성을 가질 것
 - 제공받은 데이터 + 크롤링 등으로 제3자 취득할 수 있는 데이터
2. 인사이트가 도출되었을 때, 쉽게 실행할 수 있는 액션이 떠오르는 방향일 것
3. 모두가 100% 동의할만한 인사이트는 상식 or 이미 현업에서 시도했을 가능성이 있기에 도전적인 인사이트일 것

타겟 규정

• 구매 비율 분석

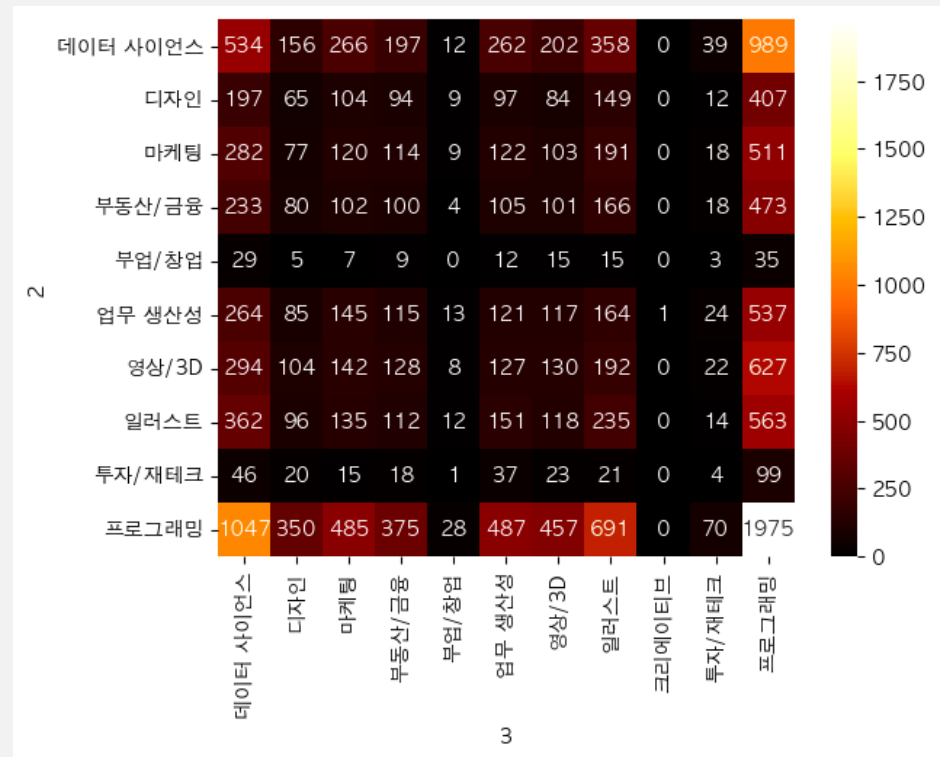
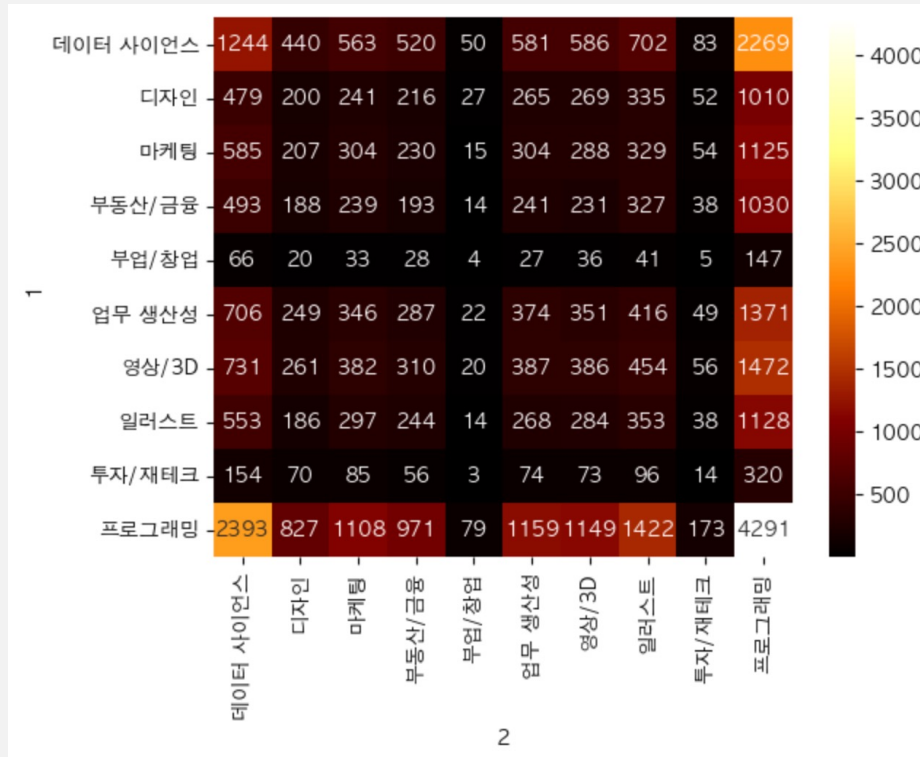
- 22년 내 3회이상 구매한 고객의 구매자 비율(24.2%)에 비행 구매 누적액 비율(43.8%)이 높음
- 위 유저를 파워유저로 규정하고, 특성을 파악해 파워유저가 될 가능성을 높이는 액션을 찾고자 함.



코호트(파워유저) 분석

• 결제차수별 카테고리 연관성 분석

- 상관관계 존재 시, 카테고리 해당 타겟에게 액션할 예정이었으나 유의미한 상관관계 없음.



코호트(파워유저) 분석

- 3회 이상 구매자와 2회 구매자의 첫결제 이후 구매간격 비교
 - 평균값에서 약 30일에 차이가 남.

df_2_pivot.describe()	
count	간격
count	24787
mean	<u>128 days 00:25:52.947714528</u>
std	91 days 01:24:11.535697943
min	0 days 00:00:04
25%	49 days 20:02:38.500000
50%	113 days 00:03:39
75%	193 days 13:07:20.500000
max	<u>362 days 12:10:27</u>

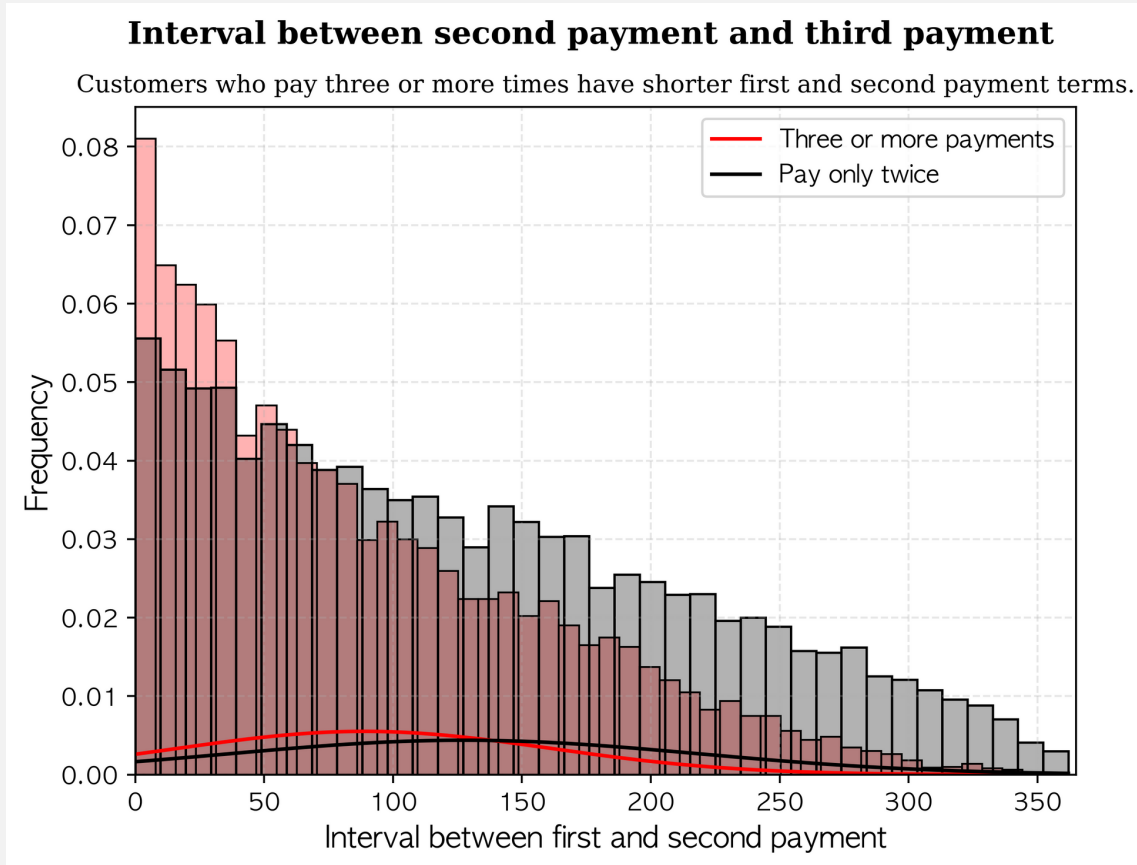
▲ 2회 구매자(3회이상 구매자 제외)

df3.describe()	
count	간격
count	18699
mean	<u>89 days 07:21:56.919086582</u>
std	72 days 09:43:40.897946427
min	0 days 00:00:02
25%	29 days 11:01:04.500000
50%	71 days 10:03:23
75%	135 days 21:34:28
max	<u>352 days 14:11:46</u>

▲ 3회이상 구매자

코호트(파워유저) 분석

• 그래프 분석



- ‘2회 구매자’와 ‘3회이상 구매자(파워유저)’로 그룹을 나누고 1차결제와 2차결제의 구매 텀 파악
- 확률밀도함수(PDF)를 통해 모수의 영향을 받지 않도록 함.
- 정규분포 그래프를 그려, 교차 지점 파악

결론

- 타겟 그룹인 **‘파워 유저’**와 2회만 구매한 그룹을 비교했을 때, 22년 첫 구매 후, 130~140일 이전에 두번째 구매가 이루어진 고객이 **‘파워 유저’**일 확률이 높았다.
- 이를 통해 23년 1회 구매 고객을 대상으로 130일 전 두번째 구매가 이루어질 수 있도록 푸쉬, 쿠폰발급 등에 액션을 하게 되면 **‘파워 유저’** 전환율을 높일 수 있다는 가설 도출
- 실행 단계에서 가설 검증하기 위해 푸쉬, 쿠폰 발급 등의 액션을 통해 130일 전 두번째 구매가 이루어진 고객들을 대상으로 **‘파워유저’**로의 전환율을 현재 데이터셋과 비교해서 의미있게 증가한다면 가설이 맞다는 결론을 낼 수 있다.

한계점

- **해당 가설을 인과관계가 아닌 상관관계일 수 있지 않나?**
 - 리서치 결과, 협업에서는 A/B 테스트 등 가설을 검증하는 액션의 결과로 인해 인과관계인 것을 밝힌다. 인과관계인지 확인하고 싶은 가설을 제시하는 것이 데이터 분석의 역할이라고 생각한다.
- **1년치 데이터만 편집된 데이터이기에 ‘파워유저’는 어쩔 수 없이 첫 구매와 두번째 구매 텀이 짧은 것이 아니냐?**
 - 해당 데이터 내에서 결론을 내다보니 어쩔 수 없는 부분이 있었음.
 - 다른 데이터도 있다면 성인교육 플랫폼의 핵심가치인 강의 수강률, 시청시간 등이 더 의미있다고 판단되어 짐.

한계점

- 구매 텀을 빠르게 한다는 것이 강의플랫폼으로 의미가 있는가? 오히려 강의를 똑바로 이용하지 못하고 있다는 것이 아닌가?
 - 발표자료에 넣지 않았지만 성인교육시장이 코로나 끝나고 축소되는 추세이며, 스타트업 투자도 이제 MAU, DAU 등 매출과 상관없는 지표로 받을 수 있는 시기가 지났다고 판단됨.
 - 다른 데이터도 활용하면 좋지만 지금 데이터에서 의미있는건 결제데이터라고 판단함.
- 원래는 액션을 ‘1번구매한 고객이 두 번째 구매를 하는 텀이 XXX일 내 한다면 **‘파워유저’** 일 확률이 YY%이다, 그래서 XXX일 내 두번째 구매가 일어날 수 있게 쿠폰, 푸쉬알림 등의 액션을 해야한다.’ 라는 최종결론을 내고 이를 시각화하고자했으나 통계학 지식 및 코딩 실력 미숙으로 인해 확률밀도함수(PDF)에서 숫자를 도출해내지 못함.

감사합니다!