SECTION 3

데이터 추출과 해석&북극성 지표 도출

SECTION 3

데이터 추출과 해석& 북극성 지표 도출

Unit

1. 데이터 추출 전 체크 포인트
2. 데이터 추출 4가지 요건
3. 데이터 해석
4. 북극성 지표 도출 & 유의점

Unit 3.1

데이터 추출 전 체크 포인트

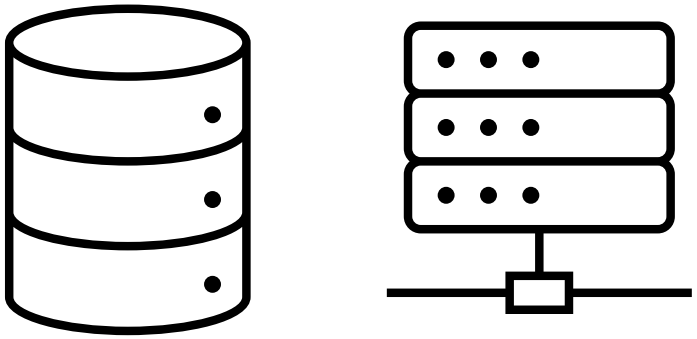
데이터 추출 요건 정리 전

데이터 이용 가능 여부 & 비중과 강도

2가지를 꼭 확인 해야 해요!

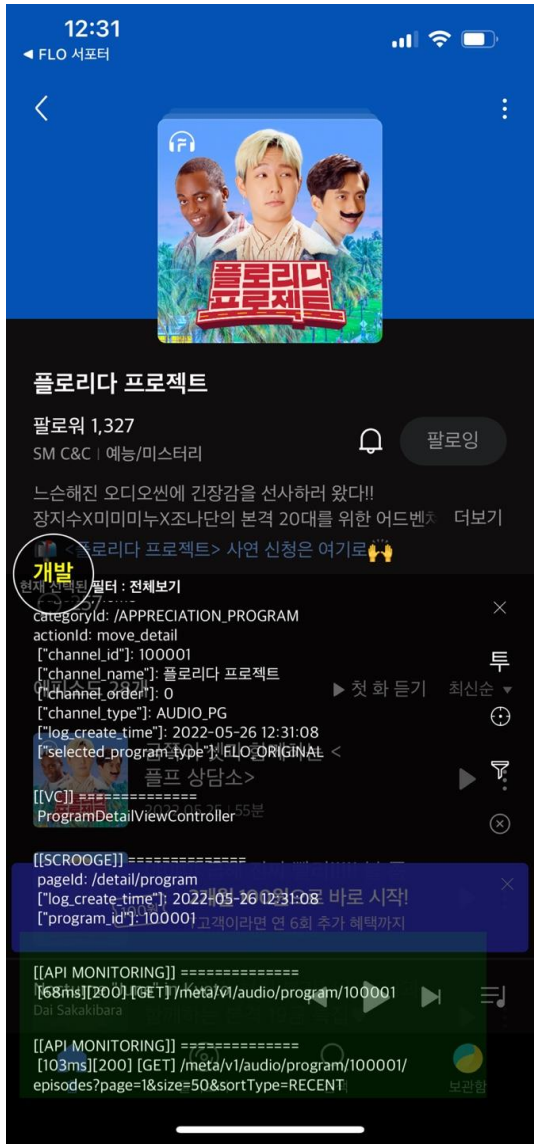
1. 데이터 이용 가능 여부

DB로 정리되어 있거나 최소한 Log가 있어야 추출 가능함



Log란?

- 앱 내에서 유저가 일으킨 행동이 **단순하게** 저장되는 것을 의미
- 로우 데이터가 계산이나 가공을 거치지 않고 있는 그대로 저장됨
- 마케팅 활동에 인사이트를 제공하는 정보와 단순 행동 데이터가 혼재함



DB란?

- 행과 열로 정보가 정리되어 있는 데이터 셋을 의미
- 하나의 DB에서 제공하는 정보가 20가지 이상을 넘어가지 않음
- 함께 있으면 더 의미를 가지는 상호 보완적인 정보끼리 모여 있음
- 반면 Log는 모든 정보가 한데 모여 있기 때문에 분석이 어려움

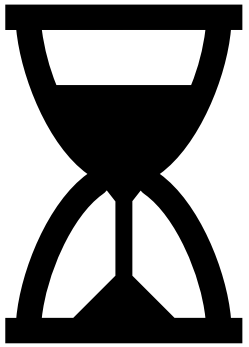
member_no	purch_date	purch_qty	purch_value	purch_item_description
jsh1234	01-01-22	1	30,000	{토마토}
fjdkjf12	01-02-22	5	56,000	{파스타링귀니면;자두200g;고랭지수박....}
jkjsnm90	01-03-22	2	20,000	{남성면도기x3;생수250gx6}
pwljlo351	01-04-22	3	90,000	{파마산치즈;....}
znljwjkk6	01-05-22	1	150,000	{페르마600n원두그라인더}
djkwhkjf	01-06-22	5	30,000	{이천쌀1kg;해태과자모음;...}

데이터 이용 가능여부 확인해야 하는 이유

- 서비스를 기획할 때 'Log를 심는' 의사결정이 필요함
- A라는 액션은 중요한 액션이니 성과측정을 위해 Log를 심자
- 주기적으로 찾아봐야 하니까 DB까지 셋업
- 이 작업이 선행되어 있지 않으면 **데이터를 추출할 수 없음**

첫번째, Log를 심는다

- 개발자와 상의해서 Log를 심고 DB까지 셋업
- 리소스를 따기 위한 설득이 필요하고
- 앱 배포 이후부터 데이터가 쌓이기 때문에 시간이 오래 걸림



두번째, 대체 데이터를 찾는다

대체할 수 있는 데이터 탐색

70%정도의정확성이면충분해요

대체 데이터 찾는 방법

넷플릭스 독점 콘텐츠 시청 횟수와 리텐션은 상관관계가 높을 것이다

자료 출처: Alliance for decision Education

가설 수립 의도: 대체제가 없는 서비스

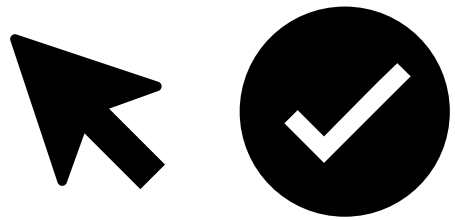
1. 주기적인 구매 니즈
2. 선택 피로도
3. 대체제가 없음

▼

넷플릭스 독점 콘텐츠 시청 횟수와 리텐션

가설 수립 의도

- 가설 수립 했던 의도를 떠올려보고
- 그 의도를 충족시키는 다른 지표를 탐색
- **시청 횟수**라는 지표의 의도는 '독점 콘텐츠를 많이 이용하는'
- 독점 콘텐츠 **플레이 버튼** 클릭 or 독점 콘텐츠 **찜하기 버튼** 클릭
- 같은 비슷한 의도의 액션으로 대체



2. 비중과 강도

팬 마케팅

2. 비중과 강도

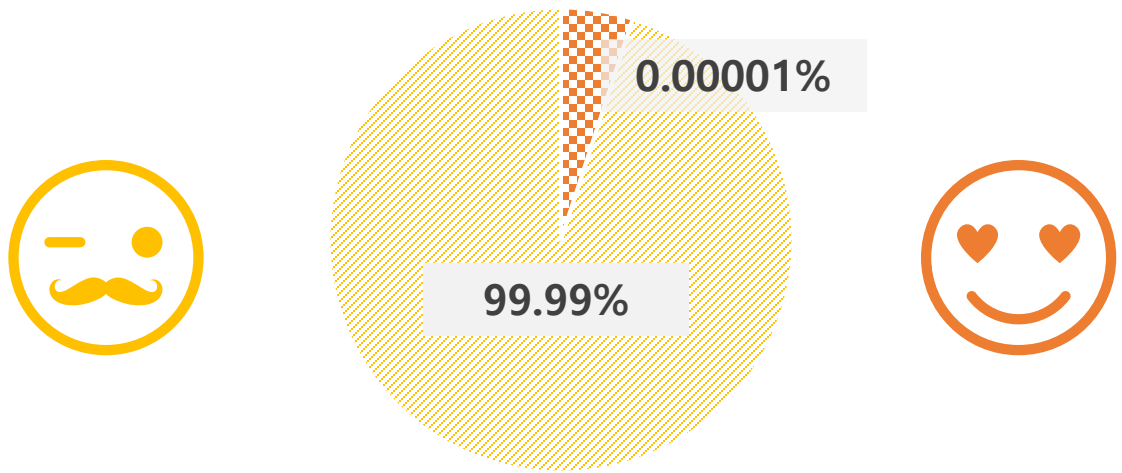
서비스가 괜찮다고 여기는 고객 100만명보다
서비스를 사랑하는 100명의 고객이 훨씬 낫습니다

by 폴그레이엄

2. 비중과 강도

0.0001%(100/100만)= **비중**

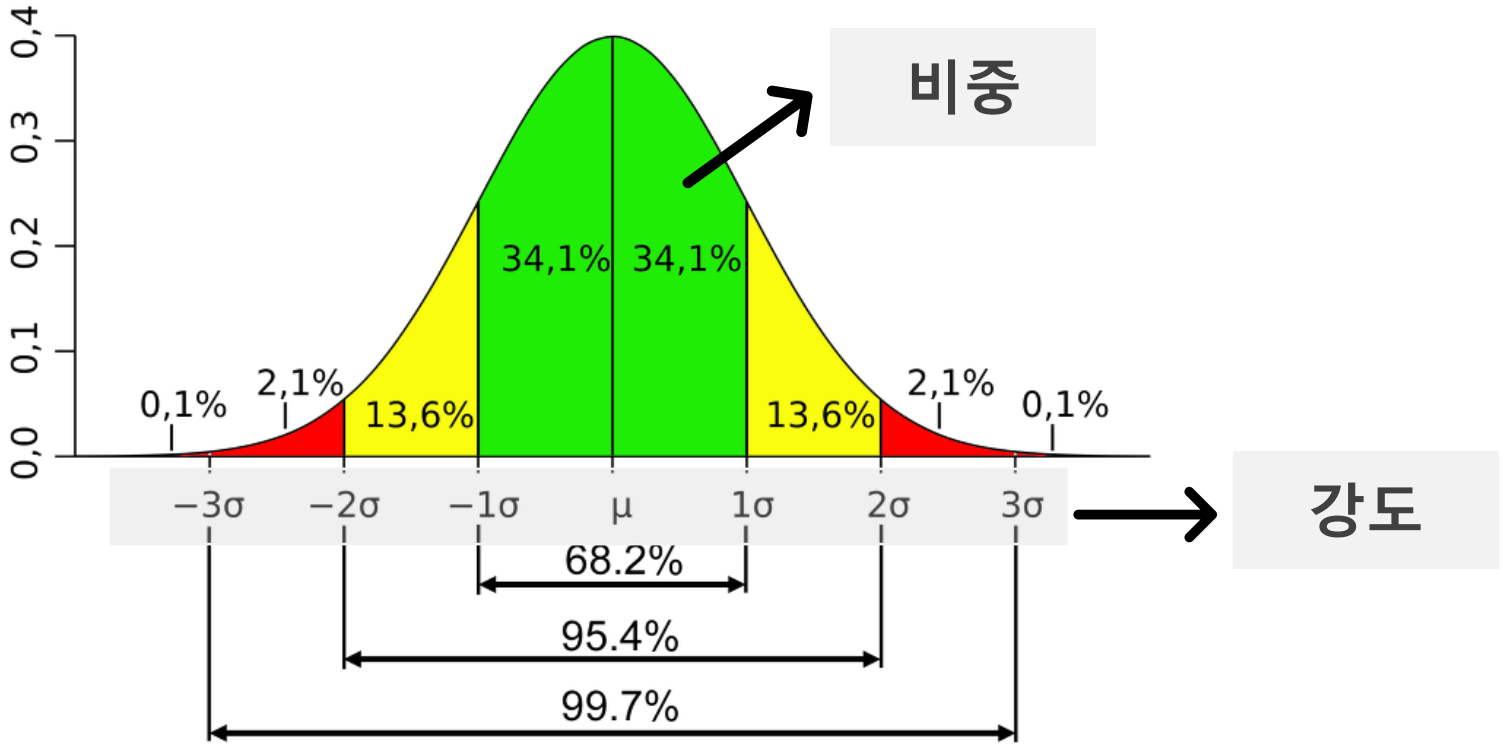
사랑하는 = 강도



액션 정의하기

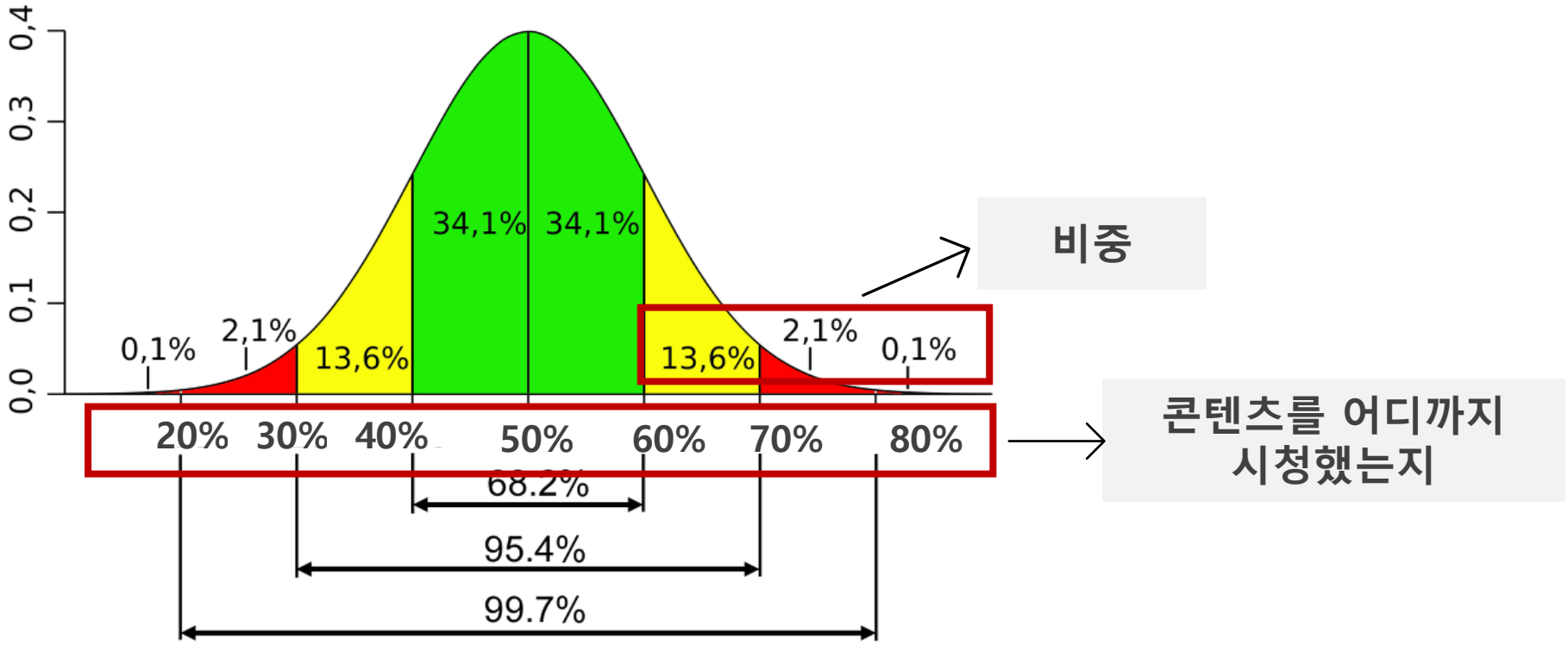
넷플릭스 독점 콘텐츠 시청 횟수와 리텐션은 상관관계가 높을 것이다

유저 행동 분포



이미지 출처: https://opinion.krsocsci.org/217

비중과 강도



이미지 출처: https://opinion.krsocsci.org/217

비중&강도 고려해야 하는 이유

- 비중과 강도가 대부분 **반비례**하기 때문
- **강도를 높게** 잡으면 서비스 애착도는 높겠지만 비중이 너무 적음
- **비중을 넓게** 잡으면 서비스 애착도가 낮은 유저들이 섞여서
- 원하는 비즈니스 임팩트를 얻기 어려움
- **적당한 수준의 비중과 강도**를 마케터가 전체 분포 데이터를 보고 판단해야 함

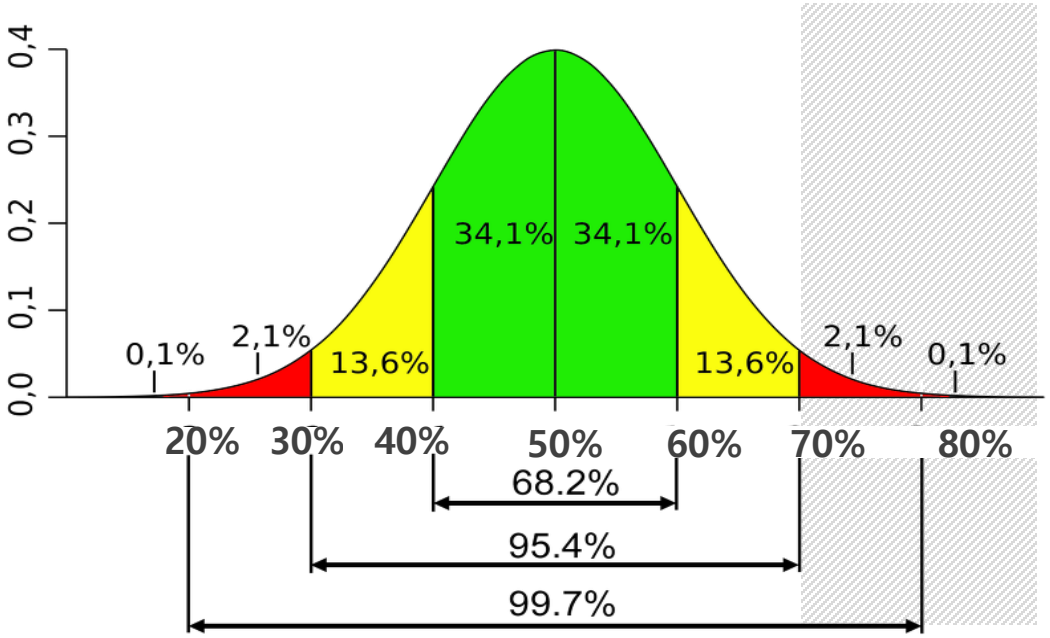
액션 정의하기

분석 목적에 맞는 비중&강도에 맞춰 의사결정

넷플릭스 독점 콘텐츠 시청 횟수

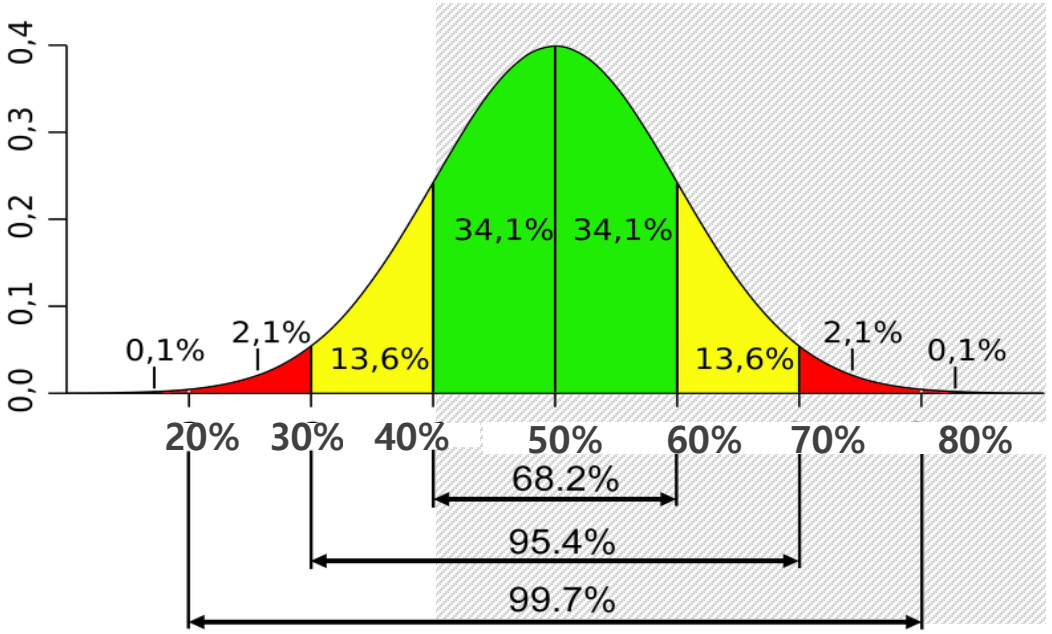
CASE 1. 비중 < 강도

- 목적이 팬을 대상으로 하는 마케팅인 경우
- 비중은 줄이더라도 강도를 높이는 방향으로 시청 액션을 정의
- 콘텐츠 전체 길이의 70%는 최소 봐야만 시청으로 인정
- 만약, 60%만 시청하고 껐다면 시청으로 불인정
- (X=70% 가 되는 구간의 면적의 합) = 비중, 전체 유저의 2.2%



CASE 2. 비중 > 강도

- 목적이 넓은 타겟을 대상으로 하는 마케팅인 경우
- 비중은 넓히고 강도를 줄이는 방향으로 시청 액션을 정의
- 전체 유저의 80%는 커버하는 시청 액션을 잡고 싶다면
- 이전 케이스와는 반대로
- 면적의 합이 80%가 되는 지점을 찾은 후 x축의 액션강도를 확인
- 40%만 콘텐츠를 보고 껐더라도 시청으로 인정
- 강도가 중요했던 CASE1번 대비 관대한 액션 기준



정리

1. 북극성 지표 가설 수립
2. 가설 검증에 필요한 데이터 추출 전 체크 포인트
3. 이용 가능여부 확인
4. 비중과 강도 참고하여 액션 정의
5. 데이터 추출 요건 정의

Unit 3.2

데이터 추출 4가지 요건

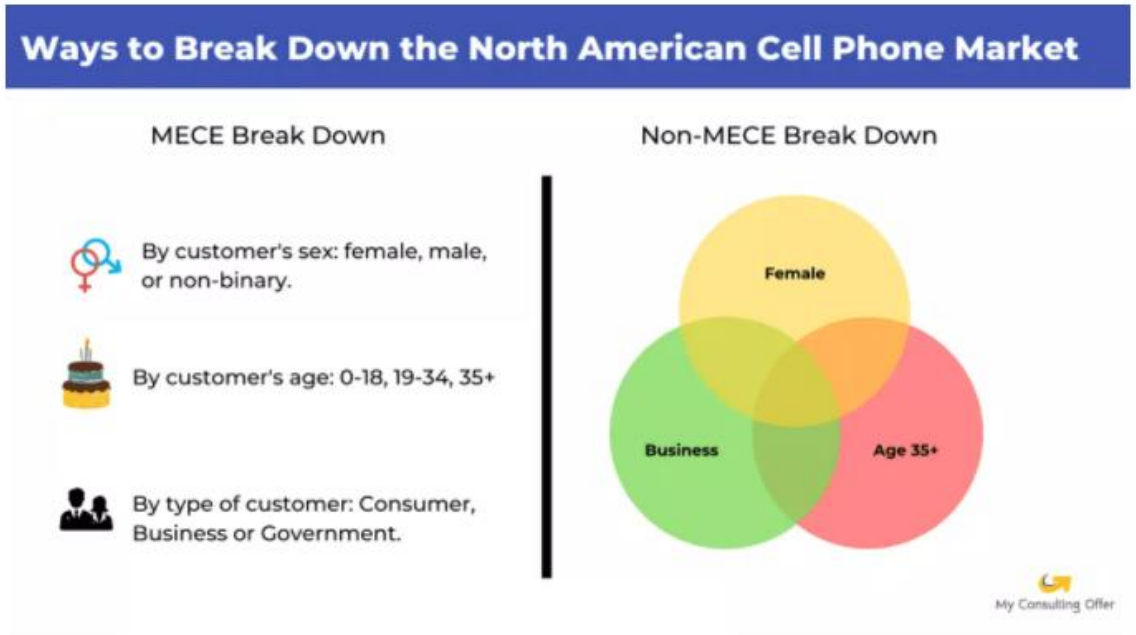
하나라도 해당된다면

- ☑ 데이터는 언제부터 언제까지 뽑아야 하나요?
- ☑ 21년 10월에 서버 이슈 있었는데 예외 처리 안 하실 건가요?
- ☑ 무료 이용권 계정도 포함해서 데이터 보실 건가요?
- ☑ 시청의 정확한 액션 정의는 어떻게 되나요?
- ☑ 리텐션의 정확한 정의가 뭔가요?

이런 말 들어요

일 진짜 성의 없이 하네 

MECE 정리법



Mutually Exclusive Collectively Exhaustive

겹치지 않으면서 빠짐없이 나눈 것

이미지 출처:https://www.myconsultingoffer.org

가설

넷플릭스 독점 콘텐츠 시청 횟수와 리텐션은 상관관계가 높을 것이다

데이터 추출 4가지 요건

기간/모수/원인지표/결과지표

데이터팀에게는 원인지표와 결과지표의 상관관계를 분석 요청하면 돼요

데이터 추출 요건 2가지

넷플릭스 독점 콘텐츠 시청 횟수와 리텐션

원인지표 결과지표

1. 기간

요건	정의 방법
시작일&종료일	데이터 추출 목적에 부합하는 기간 선정
데이터 길이	너무 짧으면 추출 목적에 부합하지 않고 너무 길면 트렌드와 다른 데이터가 포함
추출 주기	서비스 특징을 고려해서 추출 주기를 의사결정
아웃라이어	평소 앱 이용 패턴과 다른 데이터가 포함된 기간은 제외
지표	원인 &결과 지표의 기간을 각각정의

원인지표 기간

요건	정의 내용
시작일&종료일	21.07.01~21.12.31
데이터 길이	최근 6개월
추출 주기	주(Weekly), 월요일 시작
아웃라이어	제외없음(오징어 게임 포함, 비포함 2가지 CASE로 추출)

결과지표 기간

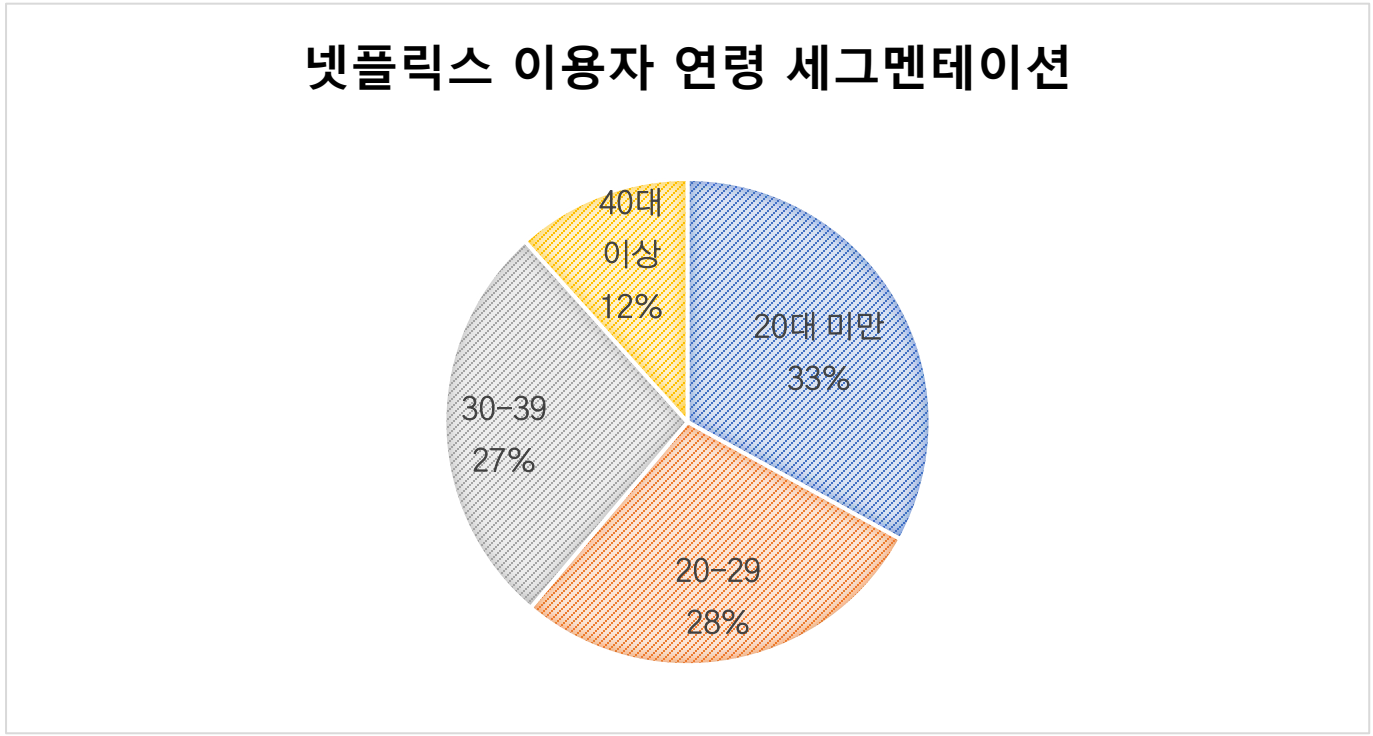
요건	정의 내용
시작일&종료일	21.07.01~21.12.31
데이터 길이	최근 6개월
추출 주기	월(Monthly), 매월 마지막 날
아웃라이어	제외없음(오징어 게임 포함, 비포함 2가지 CASE로 추출)

2. 모수

추출 대상을 누구로 할 것인지

- SEGMENTATION 하고자 하는 속성 정보 함께 추출하기

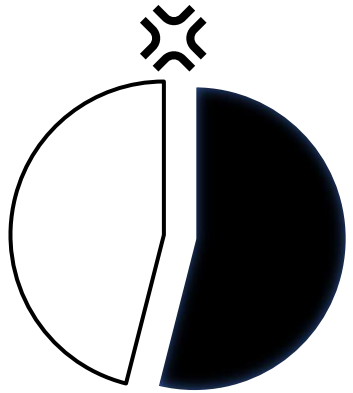
세그멘테이션



각 연령 그룹별로 독점 콘텐츠 시청 횟수와 리텐션의 상관관계를 관찰

세그멘테이션

가장 유의미한 세그멘테이션은 서비스 이용 목적



맥도날드 고객조사

밀크 쉐이크 판매 40%가 출근 시간대인 이른 아침에 이루어진다는 것을 알아냈습니다. 또한 구입하는 사람들은 드라이브 스루나 테이크 아웃을 이용했다. 밀크 쉐이크는 옷이나 운전대를 더럽히지 않고 운전에 방해가 되지 않았다. 반면 오후에는 학생이나 주부 등이 간식용으로 많이 구매했습니다.

이미지 출처:Flaticon/ 자료출처 :http://www.thefirstmedia.net

맥도날드 세그멘테이션

구매 시간대 (오전 VS 오후)

이용 목적이 달라지면 원인 지표도 달라지기 때문

모수 정의

특정일에 첫 구매가 발생한 유저

연령 정보를 함께 추출해서 세그멘테이션에 활용

3. 원인지표

넷플릭스 독점 콘텐츠 시청 횟수

- ‘시청’은 데이터 추출 전 체크 포인트2. 비중과 강도 파트 참고, 콘텐츠 50%이상 시청
  - ‘독점’ : 넷플릭스 ‘오리지널’ 콘텐츠 vs 넷플릭스’에만’ 있는 콘텐츠

CASE1. DB 있음

- DB가 정리되어 있는 경우 DB명과 칼럼 값을 정리해서 전달

- DB명 : exclusive_contents_watch_count
- Column명: count

CASE2. DB 없음 & Action Log 있음

- 기획자에게 화면정의서를 요청하거나
- 데이터 추출 의도를 설명하고 기획자 또는 개발자에게 확인 요청

- page_id='/contents_play'
- action_id='/play_view'

4. 결과지표

리텐션

‘리텐션’의 정확한 의미가 무엇인지 정의 필요

e.g, 재방문, 콘텐츠 재생, 이용권 재구매 등

데이터 추출 요건 정의

요건	정의 내용
모수	21.07.01 이용권 첫 구매가 발생한 유저 오징어 게임 시청 포함 vs 제외 2가지 case로 나뉘어서 추출
원인지표	• 주 평균 독점 콘텐츠 시청 횟수 • 독점= 넷플릭스 오리지널 콘텐츠만 의미 (e.g,cotents_type=original) • 시청 횟수= 콘텐츠의 50% 이상 시청한 경우 시청횟수 1회로 간주 • 데이터 시작일 : 2021/07/01 • 데이터 종료일 : 2021/12/31 • 데이터 추출주기 : 주(weekly)
결과지표	• 리텐션 = 이용권 재구매 • 데이터 시작일 : 2021/07/01 • 데이터 종료일 : 2021/12/31 • 데이터 추출주기 : 월(Monthly)

CASE STUDY

원거리 음식점 배달 경험 횟수와 리텐션 상관관계

생각을 정리해보세요

이미지 출처:Freepik

체크 리스트

- 데이터 이용가능 여부
  - ‘배달거리’ 데이터 이용 불가능한 경우 대체 데이터를 탐색 가정
- 데이터 비중과 강도
  - ‘원거리 배달’을 어떻게 정의하면 좋을지
- 기간
  - 분석 목적에 부합하는 너무 짧지도 길지도 않은 데이터
- 모수
  - 세그멘테이션에 필요한 속성 정보 (인구통계, 서비스 이용 목적 등)
- 원인지표
- 결과지표

데이터 이용 가능여부 확인

원거리 = 배달거리 데이터

없는 경우, 가설을 세웠던 의도를 생각해보면

대체 가능 데이터 찾기

배달예상시간

- 배달 예상 시간이 길수록 먼 것으로 간주
  - 배달료도 대체 가능

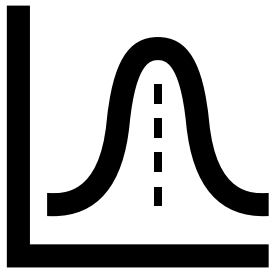
비중과 강도

원거리

몇 km를 원거리로 할 지 의사결정 필요

분포 확인

- 평균적으로 우리 유저들은 몇 km안에서 배달을 시켜 먹는 지
- 가장 먼 배달거리는 어느 정도 인지
- 전체 80% 정도 유저를 커버하는 배달 거리는 몇 km인지



데이터 기간

- 유저의 평소 이용 패턴과 다른 아웃라이어가 없는 기간
- 데이터를 관찰하기에 적당한 길이로 지정
- e.g, 대형 프로모션, 크리스마스
- 원인지표= 10월 한 달
- 결과지표= 11월 한 달

데이터 모수

- 동일한 서비스 이용 경험을 가진 유저를 분석해야 함
- 특정일에 배달거리 5km 이상 경험이 처음 발생한 유저
- 세그멘테이션 정보 e.g, 유저 주소지 정보 중 '구 ' 정보

원인지표

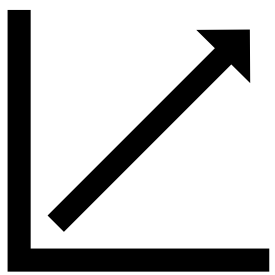
원거리 음식점 배달 경험 횟수

- 원거리 = 배달거리 5km 이상
- 배달 경험 횟수= 인당 누적 배달 경험 횟수

결과지표

리텐션

- 원거리 배달건수
  - 총주문금액
  - 총주문횟수



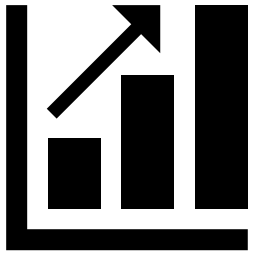
결과지표 기간 재정의

As-is

11월한달

To-be

9~11월



데이터 추출 요건

요건	정의 내용
모수	• 2021년 10월 1일 원거리 음식점 첫 배달이 발생한 사람 • 주소지(구)
원인지표	• 인당 누적 원거리 음식점 배달 횟수 • 원거리= 배달거리 5km 이상 • 데이터 시작일 : 2021년 10월 1일 • 데이터 종료일 : 2021년 10월 31일
결과지표	• 주 평균 원거리 배달 건수, 총 주문금액, 총 주문횟수 • 데이터 시작일 : 2021년 9월 1일 • 데이터 종료일 : 2021년 11월 30일 • 추출 주기: 주

CASE STUDY

데이터 추출요건 정의

가설

토스 큐레이션 페이지 이용 경험과 리텐션

실습
데이터 추출요건 정의

체크리스트

- 데이터 이용가능 여부
- 데이터 비중과 강도 참고해서 액션 정의
- 데이터 추출 기간
- 데이터 모수
- 원인지표
- 결과지표

실습
데이터 추출요건 정의

생각 포인트

- 데이터 이용가능 여부
  - 가설 검증을 위해 어떤 데이터를 봐야 할까요?
- 데이터 비중과 강도
  - 큐레이션 이용 경험의 **'이용'**을 어떻게 정의해야 할까요?
- 데이터 추출 기간
  - **추출 기간**과 **주기**는 어느 정도로 정의하면 좋을까요?
  - 평소와 다른 행동 데이터가 포함된 기간은 없을지 생각해봐요.
- 데이터 모수
  - 동일한 앱 이용 컨디션을 가지는 모수 조건을 생각해봐요.
- 원인지표
  - **큐레이션 페이지**를 데이터로 어떻게 정의해야 할까요? 가설 의도가 가장 잘 반영된 정의를 생각해봐요.
- 결과지표
  - 리텐션을 어떻게 정의해야 할까요? 가설 의도가 가장 잘 반영된 정의를 생각해봐요.

데이터 이용 가능여부 확인

큐레이션 페이지 이용

페이지 체류시간

스크롤 n% 이상

비중과 강도

페이지 체류시간

얼마나 체류하면 큐레이션 페이지를 '이용'했다고 볼 건지 정의

데이터 기간

- 큐레이션 페이지 런칭 시점을 고려해서
- 데이터를 관찰하기에 적당한 길이로 지정
- 데이터 추출 기간 동안 동일한 조건이 맞는 지 더블 체크
- 체크 없이 전체 데이터 기간을 뽑으면 데이터 왜곡의 소지가 있음
- E.g, 큐레이션 페이지 유입을 위한 진입점 변경 + 푸시 메시지 발송
- 데이터 추출 주기 역시 유저의 접속 빈도를 고려해서 마케터가 정의할 것

원인지표

토스 큐레이션 페이지 이용 경험과 리텐션

- 큐레이션 페이지 = ‘page_id=curation’
- 큐레이션 페이지를 어디까지 포함 /제외 시킬지

Unit 3.3

데이터 해석

분석 vs 해석

데이터 단순 읽기 vs 유저 행동 이해하기

해석연습1

넷플릭스 접속횟수와 리텐션 상관관계가 없을 때

- 접속횟수 보다는 콘텐츠 만족도에 리텐션이 올라가는 걸까?
- 그렇다면 콘텐츠 만족도는 데이터로 어떻게 측정할 수 있을까?
- 콘텐츠 만족도가 높았던 콘텐츠의 공통점은 무엇이었나?
- 콘텐츠 만족도가 낮았던 콘텐츠를 많이 보면 리텐션이 낮아졌나?

해석연습2

넷플릭스 독점 콘텐츠 시청횟수와 리텐션 X

추가적으로어떤질문을해볼수있을까요?

추가적인 질문

- 왜 독점 콘텐츠 시청과 리텐션의 상관관계가 높지 않을까?
- 독점 콘텐츠 시청 '횟수'가 아니라 시청 '만족도'와 리텐션은 상관관계가 있을까?
- '시청'의 기준을 콘텐츠 100% 시청한 경우로 바꿀 경우 결과는 어떻게 변화할까?
- 유저가 독점 콘텐츠에 노출되는 퍼널은 어떻게 되나?

해석연습3

원거리 음식점 배달 횟수와 리텐션 X

추가적인 질문을생각해봐요

유저 행동 이해하기

왜 배달거리가 멀어지면 리텐션이 낮아질까?

이유를 먼저 생각해봐요.

유저 행동 이해하기

배달거리가 멀어질 수록 음식맛이 떨어진다?

마케터는 이 가설을 어떻게 디벨롭 시켜보면 좋을까요?

가설 디벨롭

- 만족도 하락을 데이터로 측정하는 방법 생각해보기
- 원거리 배달 경험 유저의 **평균 평점, 후기 내용 확인**
- 가설을 증명해 봐야겠다는 생각이 들었다면 **설문조사**
- 불만족의 원인은 정량 데이터로 파악하기 어려움
- 불만족의 원인, 만족 포인트, 개선 포인트 등을 조사
- 서비스 개선에 반영 or 마케팅 메시지로 활용

유저 행동 이해하기2

모수 정의에 '주거구' 정보 추가

'구' 정보로 데이터를 쪼개 봤을 때 상관계수 살펴보기

유저 행동 이해하기2

구조화

전체 '원거리 배달' 유저를 거주지 '구'로 나누어서 쪼개 보기

유저 행동 이해하기2

‘주거구’ 별로 상관관계가 모두 동일한지 or 다른지

- 상관계수가 높은 구 공통점
- 상관계수가 낮은 구 공통점

유저 행동 이해하기2

상관계수가 높은 구의 공통점

마케터의 직관 + 객관적 데이터(e.g, 통계청 인구정보 혹은 평균소득, 부동산 가격) back up

상관계수

	program_uniq_cnt	episode_uniq_cnt	follow_count
program_uniq_cnt	100%		
episode_uniq_cnt	44%	100%	
follow_count	24%	23%	100%
listen_days	13%	18%	4%
re_listen	-23%	-35%	-14%
Frequecny	31%	61%	16%

엑셀> 데이터> 데이터분석> 상관분석

이미지 출처: Alliance for decision Education

Unit 3.4

북극성 지표 도출 & 유의점

마케터의 통찰력

상관관계가 높은 지표 중 인과 관계 판단

- 상관계수가 높다고 인과 관계 보증 NO

상관관계 분석결과

1. 선택의 피로도가 존재하는 서비스
평점을 n번 이상 입력한 것과 리텐션은
상관관계가 높을 것이다

상관계수= 0.75

2. 대체제가 없는 서비스
넷플릭스 독점 콘텐츠 시청횟수와 리텐션은
상관관계가 높을 것이다

상관계수= 0.67

북극성 지표를 위한 질문

1. 어떤 지표를 북극성 지표로 해야 할까요?
2. 두 지표 모두 상관관계가 있기 때문에 두 지표 다 관리해야 할까요?
3. 상관계수가 높은 지표를 북극성 지표로 정의하면 될까요?

인과관계 3조건

- [x] 선행지표
- [x] 원인지표
- [x] 독립지표

1. 독립지표 아닌 경우

- 가설을 여러 개 수립했다면
- 상관관계가 검증되는 가설이 한 개 이상 나올 확률이 높음
- 원인지표들은 서로 독립적이기보다 **상호 영향을 주고 받음**
- 하나의 지표가 변화하면 다른 지표도 변화
- 각 지표들의 변화가 종합적으로 성과에 영향을 미침

마케터는..

이미지 출처: unsplash

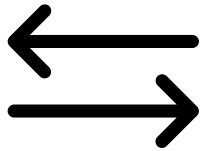
마케터의 흔한 고민들

- 모든 지표를 챙겨봐야 한다는 착각
- 북극성 지표를 발굴한다는 건
- 가장 인과관계가 있을 법한 **하나의 지표를 선택**하는 과정
- 지표가 여러 개면 집중력이 흩어지고 개선 가능성은 낮아짐

이미지 출처: freeiconslibrary

1. 지표간 영향 이해하기

넷플릭스 독점 콘텐츠 시청횟수와
리텐션은 상관관계가 높을 것이다



평점을 n번 이상 입력한 것과
리텐션은 상관관계가 높을 것이다

2. 상관관계가 인과관계 보증하지 않음

평점을 n번 이상 입력한 것과 리텐션은
상관관계가 높을 것이다

상관계수= 0.75

넷플릭스 독점 콘텐츠 시청횟수와 리텐션은
상관관계가 높을 것이다

상관계수= 0.67

인과관계 Check list

☑ 역의 인과관계
☑ 제3의 변수
☑ 우연의 일치

자료 출처: Alliance for decision Education

3. 역의 상관관계

평점을 n번 이상 입력한 것과 리텐션은
상관관계가 높을 것이다

리텐션이 높은 사람이
평점 입력하기를 좋아하는 것 아닐까?

4. 제 3의 변수

평점을 n번 이상 입력한 것과 리텐션은
상관관계가 높을 것이다

콘텐츠에 대한 만족도