



EDA 방법론

22.07.19/ DSL 7기 김한빈

1. 프로젝트 설명

- 기초 세션 설명

2. EDA란?

- EDA란?
- EDA목표 & 과정
- EDA vs CDA

3. 분석이란?

- 분석의 정의
- 분석 범위
- 분석 단계

4. 데이터의 특성

- 수치적 특성
- 수집 방식별 특성

5. 데이터 시각화

- 시각화의 목적
- 그래프의 종류
- 시각화 고려사항

6. EDA 실습 Tip!

- 데이터별 아이디어이션
- 프로젝트 예시
- 프로젝트 발표 예시



1. 프로젝트 설명

1. 프로젝트 설명

1. 기초 세션 설명

Numpy



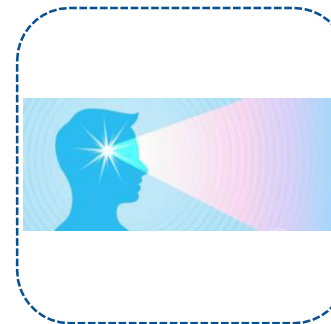
Pandas



Crawling



Visualization



Git



데이터 확인

데이터 수집

데이터 시각화

협업/아카이브

1. 프로젝트 설명

YONSEI Data Science Lab | DSL

1. 기초 세션 설명

시작 단계



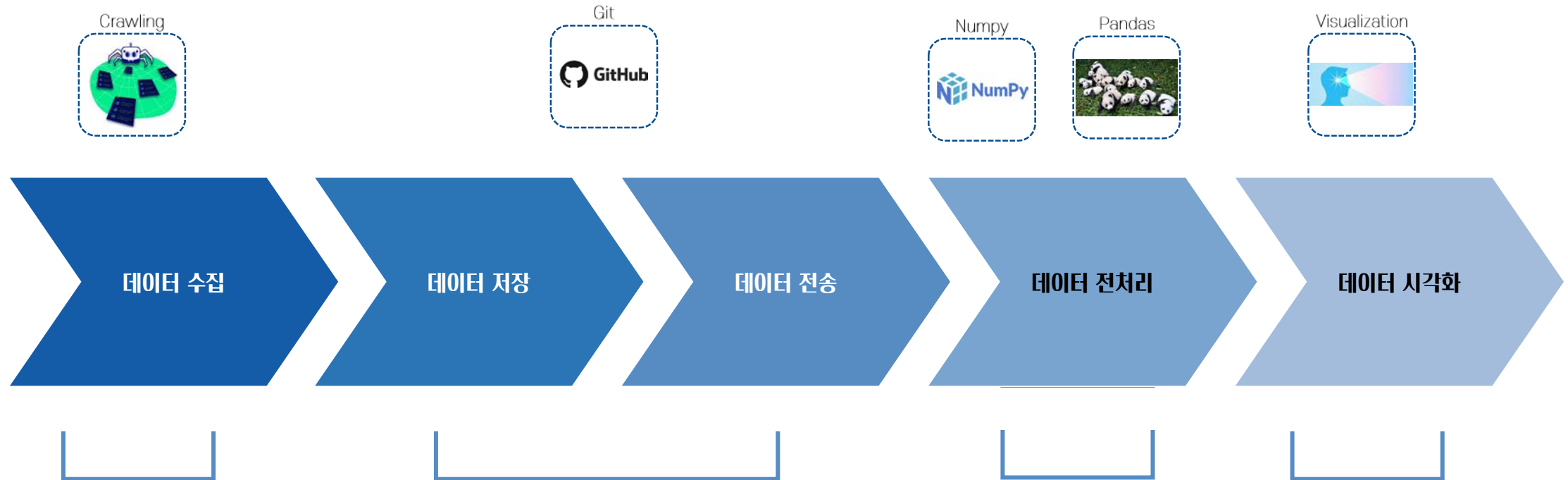
활용 단계



1. 프로젝트 설명

YONSEI Data Science Lab | DSL

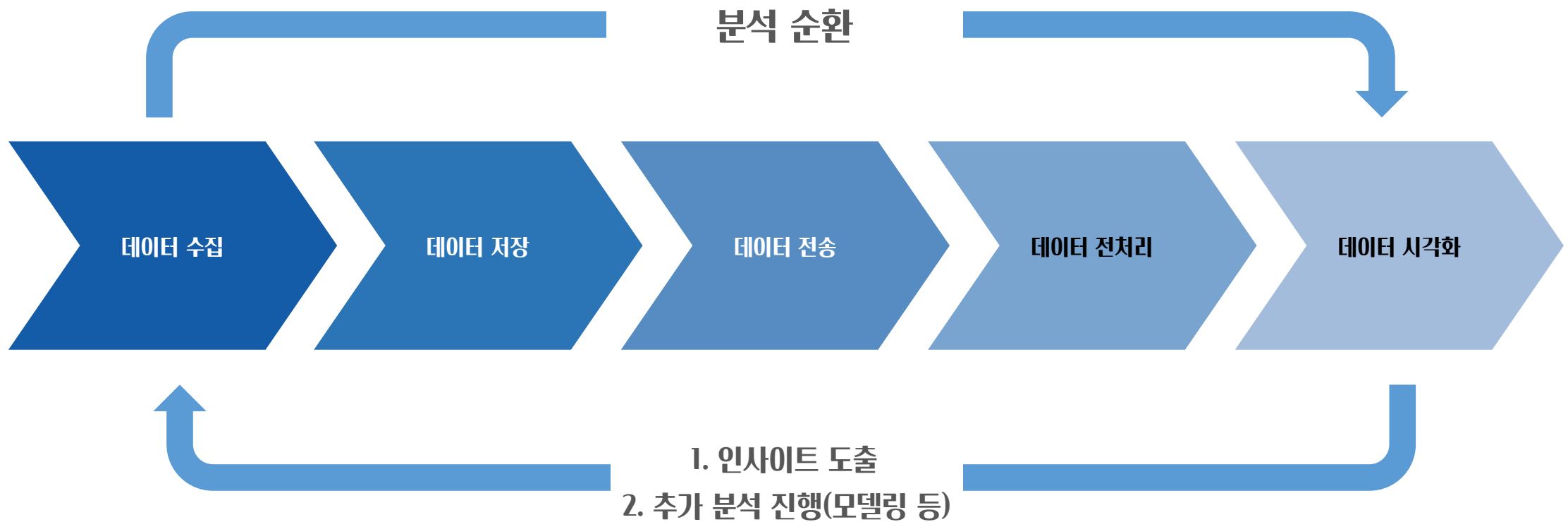
1. 기초 세션 설명



1. 프로젝트 설명

YONSEI Data Science Lab | DSL

1. 기초 세션 설명

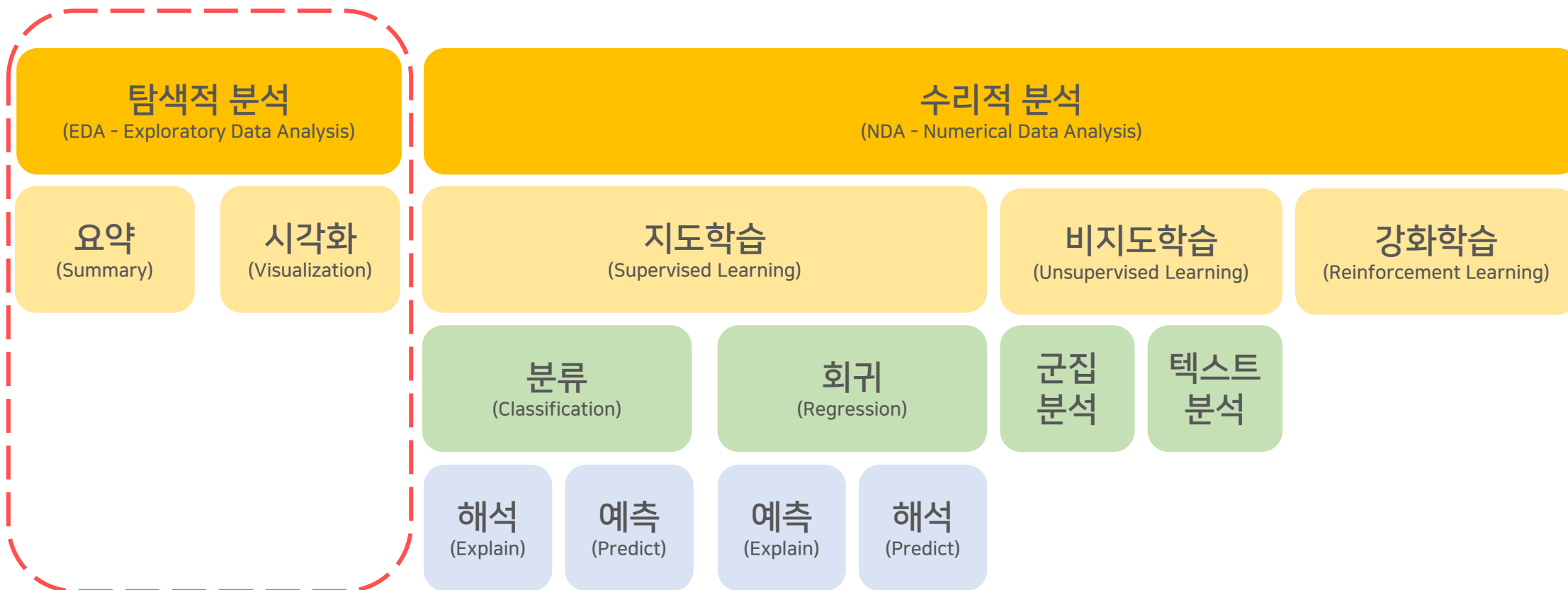




2. EDA란?

2. EDA란?

1. 데이터 분석 구조



2. EDA란?

2-1. EDA란?

Exploratory Data Analysis, ‘탐색적 자료 분석’

- 시각화, 통계 분석 등을 이용하여 데이터의 패턴을 발견하고 의미 있는 정보를 도출해내는 작업

2. EDA란?

YONSEI Data Science Lab | DSL

2-2. EDA 프로젝트 목적

데이터로부터 **유의미한 인사이트**를 도출하는 프로젝트

2. EDA란?

2-2. EDA 프로젝트 목적

데이터로부터 유의미한 인사이트를 도출하는 프로젝트

- 신뢰할 수 있다.
- 목적에 부합한다.
- 목표에 align되어 있다.
- 임팩트가 크다.
- 중요도가 높다.
- 실현 가능하다.

- 문제에서 기반 했다.
- 당연하지 않다.
- 당연하지만 인지하지 못하고 있었다.
- 다음 Action Plan이 나온다.
- **Bad 예시1** : 남자는 힘이 세다.
- **Bad 예시2** : 돈이 많은 사람이 비싼 집에 산다.

2-3. EDA가 필요한 이유

1. 초기 분석(Initial data Analysis)의 용도로 사용됨 - 데이터 구조 & 내용 파악
2. 자료 내 변수들의 분포 & 관계를 파악해 모델링을 위한 인사이트를 확보함
3. 데이터에 적용 가능한 방법론들을 탐색하고, 가정을 확인하며 데이터 품질을 관리함
4. 데이터 분석의 첫 순서이다. (first-step)

2. EDA란?

2-3. EDA의 구성

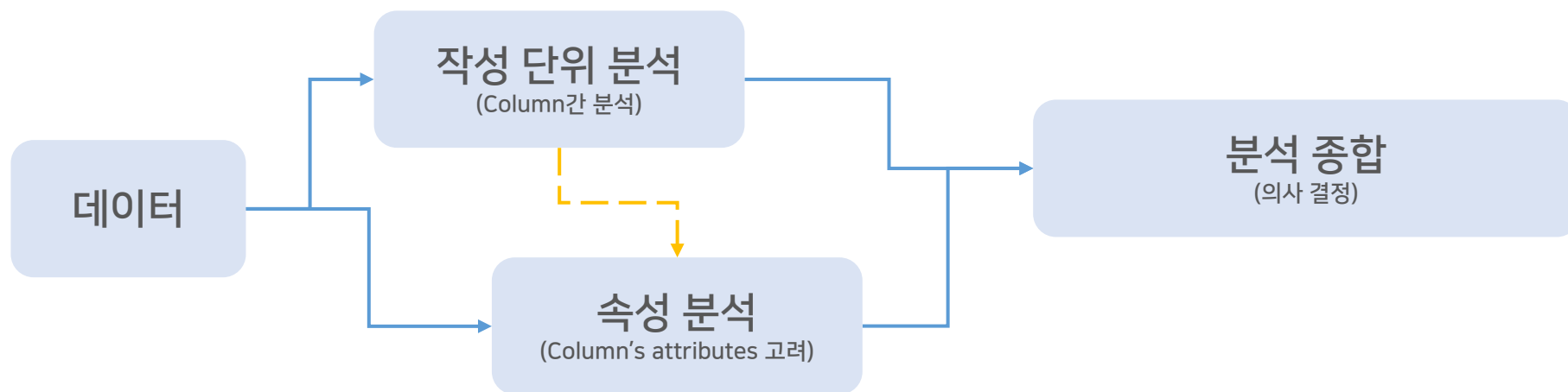
1. 기술 통계

- **수치적 자료 요약** (중심 위치 척도, 변동성 척도, 연관성 척도…)
- **표 & 그래프 시각화** (표 요약, 명목형 자료, 자료 개형, 시계열 & 변수 관계…)

2. EDA 실전

- **결측치 처리** (단위 무응답<all>, 항목 무응답<any>)
- **이상치 탐색** (표 요약, 명목형 자료, 자료 개형, 시계열 & 변수 관계…)

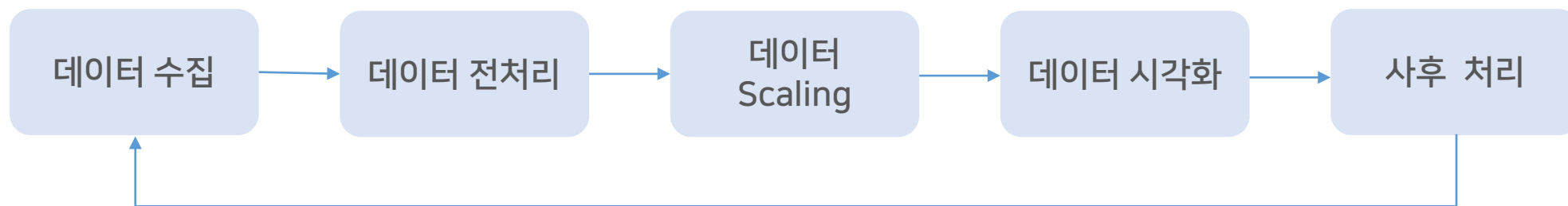
2-4. EDA 과정 - 이론적 접근



1. **작성 단위 분석:** Column간 상관관계, 분포, 작성 시 편의에 의한 Section Bias, 데이터 계층 등 확인
2. **속성 분석:** 작성 단위의 속성(attributes), 수치, 표, 기술통계량 고려
 1. 종횡비, 결측치, 이상치, 측정 오차 고려

2. EDA란?

2-4. EDA 과정 - 실전적 접근



1. **데이터 수집**: 데이터 수집 파이프라인 생성, 수집 필요 데이터 목록화
2. **데이터 전처리**: 결측치 처리, 이상치 탐색, 데이터 라벨링, 정규식을 통한 데이터 정리 etc..
3. **데이터 Scaling**: 데이터 범위 조율, 표준화/정규화, 데이터 양 조절(오버 샘플링/언더 샘플링)
4. **데이터 시각화**: 데이터 시각화 (모델링)
5. **사후 처리**: 결측치 처리, 이상치 탐색, FineTuning, 추후 분석 방향 결정

2. EDA란?

4. EDA vs CDA

EDA

Exploratory Data Analysis

데이터를 우선 살펴보면서 인사이트를 도출한다

인사이트를 얻겠다는 선입견 없이 데이터를 유연하게 탐색해본다

명확한 분석 목표가 없기 때문에 분석 결과를 보고서 아무런 인사이트도 얻지 못한 채 방황할 가능성이 높다

CDA

Confirmatory Data Analysis

가설을 세운 후에 이를 데이터를 통해 검증한다

내가 이미 가지고 있는 인사이트 가설을 검증하겠다는 명확한 목표가 존재한다

이미 어떠한 결과가 나올 것이라는 가설을 기준으로 진행하므로 가설의 검증 결과에만 집중하여 유의미한 인사이트를 주는 지표를 놓칠 수 있다

2. EDA란?

3. EDA vs CDA

EDA

Exploratory Data Analysis

데이터를 우선 살펴보면서 인사이트를 도출한다

인사이트를 얻겠다는 선입견 없이 데이터를 유연하게 탐색해본다

명확한 분석 목표가 없기 때문에 분석 결과를 보고서 아무런 인사이트도 얻지 못한 채 방황할 가능성이 높다

CDA

Confirmatory Data Analysis

가설을 세우 후에 이를 데이터를 통해 검증한다

내가 이미 가지고 있는 인사이트 가설을 검증하겠다는 명확한 목표가 존재한다

이미 어떠한 결과가 나올 것이라는 가설을 기준으로 진행하므로 가설의 검증 결과에만 집중하여 유의미한 인사이트를 주는 지표를 놓칠 수 있다

하지만 결국 **DA(Data Analysis)!**

화려함에 속지 말고 ‘분석’ 자체에 집중하자!



3. 분석이란?

3. 분석이란?

1. 분석의 정의

‘**분석**’ (分析), ‘**나**를 **분**’, ‘**조각** **석**’

- 복잡하고 많은 내용을 가진 대상을 정확히 이해하기 위해 단순한 요소로 나누어 생각함

3. 분석이란?

1. 분석의 정의

‘분석’ (分析), ‘나눌 분’, ‘쪼갠 석’

- 복잡하고 많은 내용을 가진 대상을 정확히 이해하기 위해 단순한 요소로 나누어 생각함

-> 결국 **얼마나 자세히/적절히 ‘암묵적 요소’를 ‘명목적 요소’로 나누어 사고하는지가** 분석의 품질을 결정함

3. 분석이란?

2. 분석의 목표

‘분석’은 그 목적에 따라 일정한 관점에서 해야 한다.

Why? 그래야 적절하게 중요한 요소 단위로 나누어 결론을 내고, Action Plan이 나오기 때문

-> So, 가장 먼저 **목표를 제대로 설정**하는 것이 가장 중요

3. 분석이란?

3. 분석 범위

1. 데이터 분석

- 데이터의 상태
- 데이터의 구성
- 표본의 특성

2. 공정 분석

- 분석시 선택한 방식
- 전처리 과정 (결측치 처리, 스케일링, 표준화, 필터링)
- 데이터 결합 (외부 데이터 이용)
- 표본 선택이 제대로 되었는지?

3. 분석이란?

3. 분석 범위

3. 결과 분석

- 분석 템플릿 이용 (육하원칙 등 - 해석의 객관화를 위함)
- 도메인 지식 결합 (자의적 해석, 비전문적 해석 배제)
- 목적 중심 해석

1. 분석 목적 Remind

2. 결과 제시

3. Action Plan 제시 (EDA에서는 보통 여기까지)

4. Action Plan들간의 비교 & 우선순위 제시

5. 의사결정

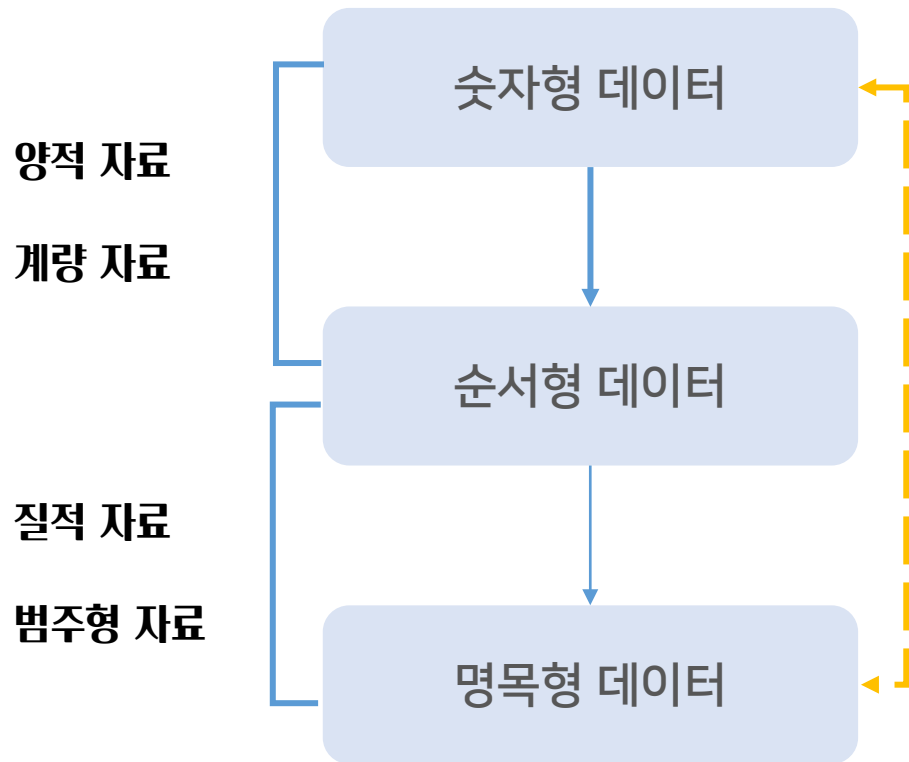
6. 사후 의사결정 결과 확인 재분석 시작



4. 테이터의 특성?

4. 데이터의 특성

1. 수치적 특성



- **숫자형 데이터:** 크기를 파악할 수 있는 숫자형 자료
 - 키(cm), 몸무게(kg), 주가(₩)
- **명목형 데이터:** 숫자가 아닌 String값 등으로 표현되는 자료, 순서/크기 파악 불가
 - 키(cm), 몸무게(kg), 주가(₩)
- **순서형 데이터:** 이름이나 문자 등으로 표현되지만, 순서를 파악할 수 있는 자료
 - 평점, 학점

4. 데이터의 특성

1. 수치적 특성

- 수치형 데이터(Continuous): 1, 2, 3, 4, 5 ... 98, 99, 100
- 수치형 데이터(Discrete): 100, 393, 45, 2222, ... 12, 885
- 범주형 데이터(Binary): '남자', '여자' / 0, X / 0, 1 ...
- 범주형 데이터(MultiClass): A+, A0, A-, B+ ... , D0, D- , F

4. 데이터의 특성

1. 수치적 특성 - 그래프

수치형 데이터 - 일변량	- 요약통계량: Mean, Max-min, Median, Standard deviation 등 - 그래프: Histogram, Boxplot 등
범주형 데이터 - 일변량	- 그래프: Bar plot, Pie plot 등
수치형 데이터 - 다변량	- 요약통계량: Correlation 등 - 그래프: Scatter plot
범주형 데이터 - 다변량	- 그래프: Side-by-side box plot 등
비정형 데이터	- 적합한 분석 모델 필요

단순	Simple	독립변수(X)가 1개
다중	Multiple	독립변수(X)가 2개 이상
일변량	Univariate	종속변수(Y)가 1개
이변량	Bivariate	종속변수(Y)가 2개
다변량	Multivariate	종속변수(Y)가 2개 이상

4. 데이터의 특성

2. 수집 방식별 특성

1. 행동 데이터(Action Data)

- 특정 대상의 행동을 지속적으로 트래킹한 데이터 (Log Data)
- 시계열 데이터 (Timely & Accumulate)

2. 상태 데이터(Status Data / Property Data)

- 대부분의 데이터, 현재 상태 값을 저장
- 특정 대상의 상태값 저장 (Linked)
- 비주기적 데이터 (Untimely & Update)

3. 응답 데이터(Response Data)

- 특정 대상의 응답을 기록한 데이터 (ex. 설문조사)
- 수집 목적이 Strict한 경우가 많고, Column별 연관성이 높다.
- 응답 내용이 인식과 관련된 경우 실제와 다르게 응답할 수 있다. (부정확 & 신뢰도 문제)



5. 데이터 시각화?

1. 시각화 정의

‘**시각화**’ (分析),

: 데이터 분석 결과를 **쉽게 이해**할 수 있도록 시각적으로 **표현**하고, **전달**하는 과정

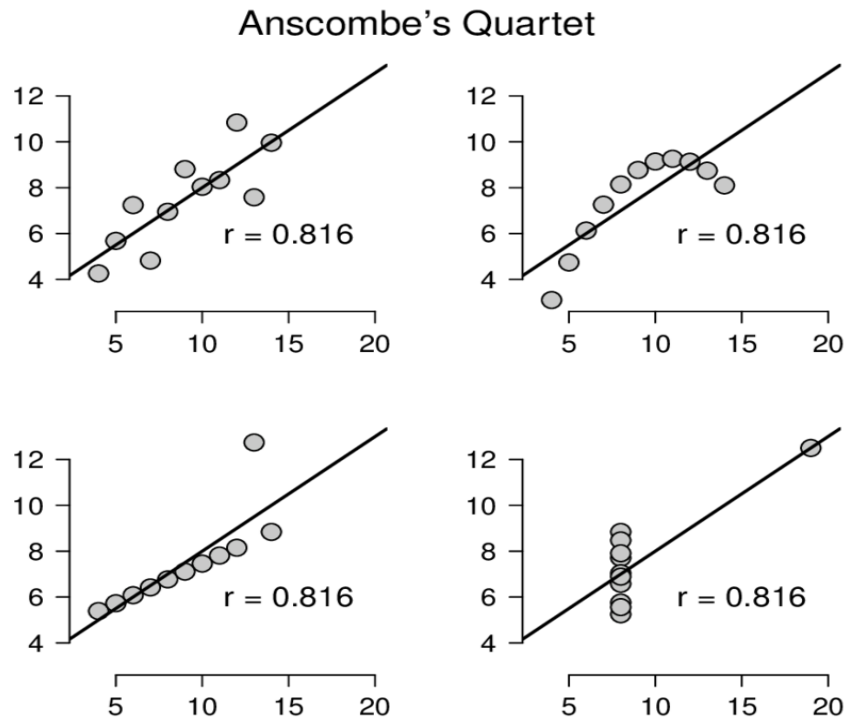
5. 데이터 시각화

2. 시각화 목표 & 고려점

- 목표: 분석자/피전달자가 쉽게 이해할 수 있어야 한다.
- 스토리텔링:
 1. 컨텍스트 설명 <- 분석의 배경 & 목적 설명
 2. 핵심 메시지 (문제 상황 & 기회) <- 시각화를 통해 설명
 3. 추천하는 액션 <- 인사이트 & To-Do

3. 그래프 종류 - 기술 통계

시각화의 중요성



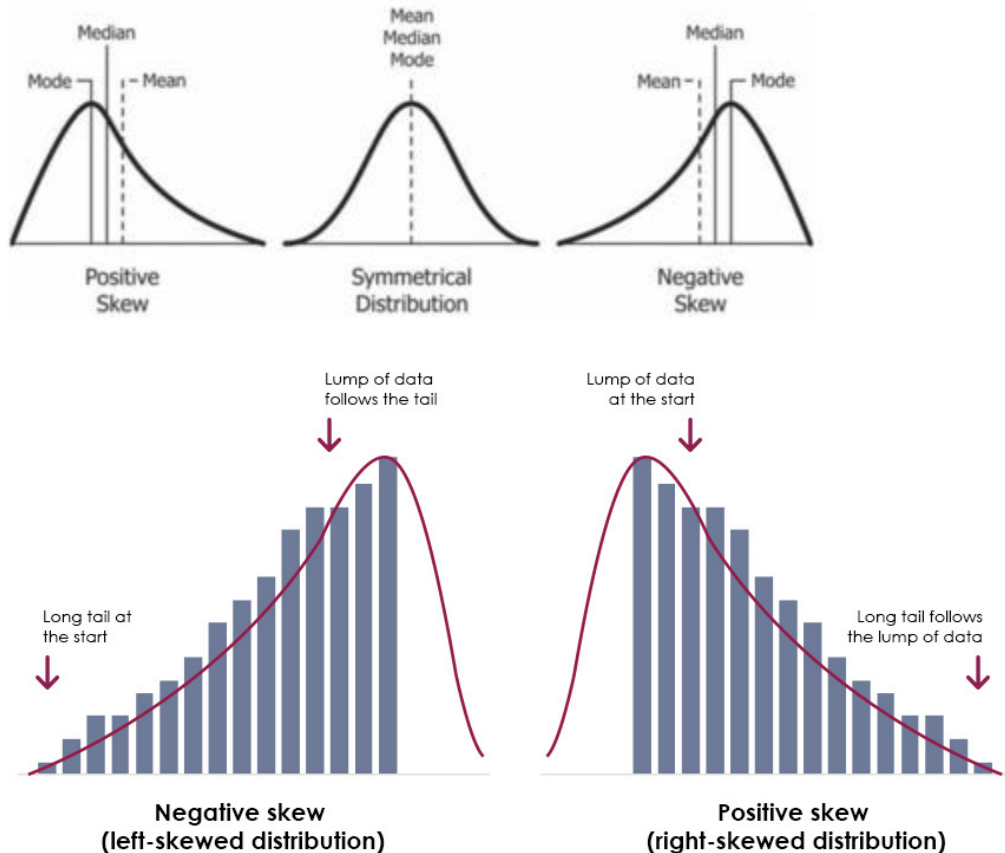
시각화의 중요성

- **Anscombe의 사인방(Quartet)**
 - 모두 같은 Pearson 상관계수를 갖지만 데이터의 분포는 천차만별이다.
- 통계량과 같은 수치에만 의존하는 것이 얼마나 위험한지 보여줌

5. 데이터 시각화

3. 그래프 종류 - 기술 통계

중심 위치 척도



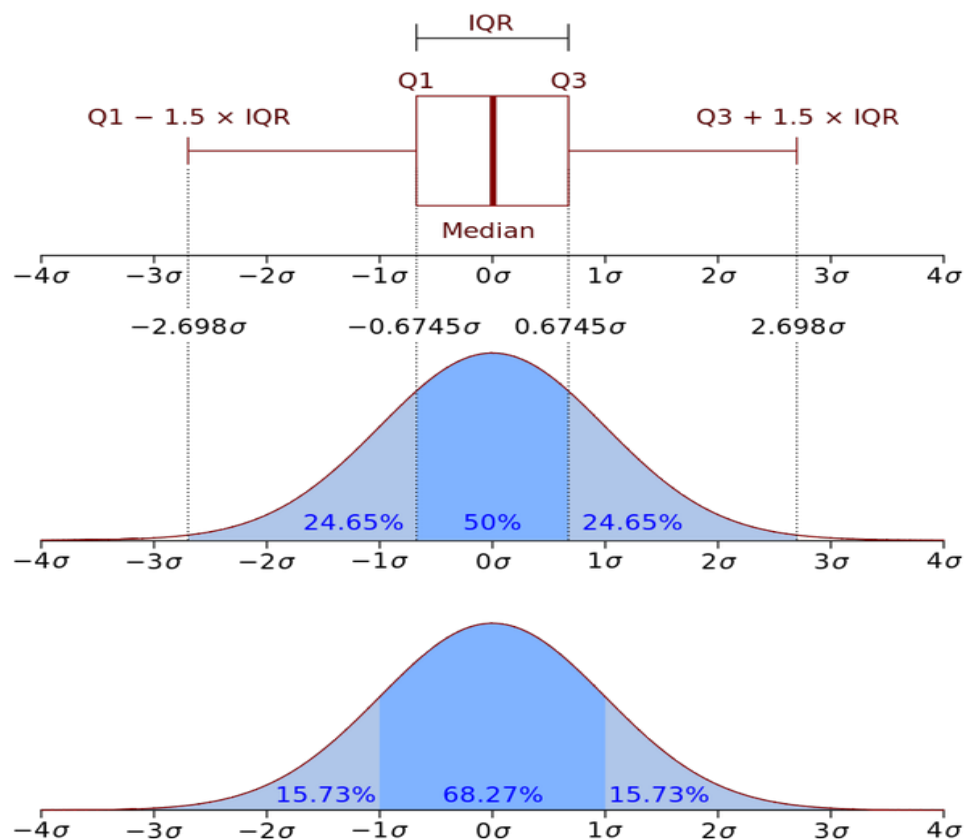
1. 중심 위치 척도

- **평균(Mean):** 모든 자료의 값을 더한 후 전체 개수로 나눈 값 → 이상치/측정오차에 민감
- **중앙값(Median):** 자료를 크기 순으로 늘어놓았을 때 가운데에 해당하는 값 → 이상치/측정오차에 덜 민감
- **최빈값(Mode):** 자료 중에서 그 빈도수가 최대인 값 → 최빈값은 여러 개 나올 수 있다

5. 데이터 시각화

3. 그래프 종류 - 기술 통계

변동성 척도



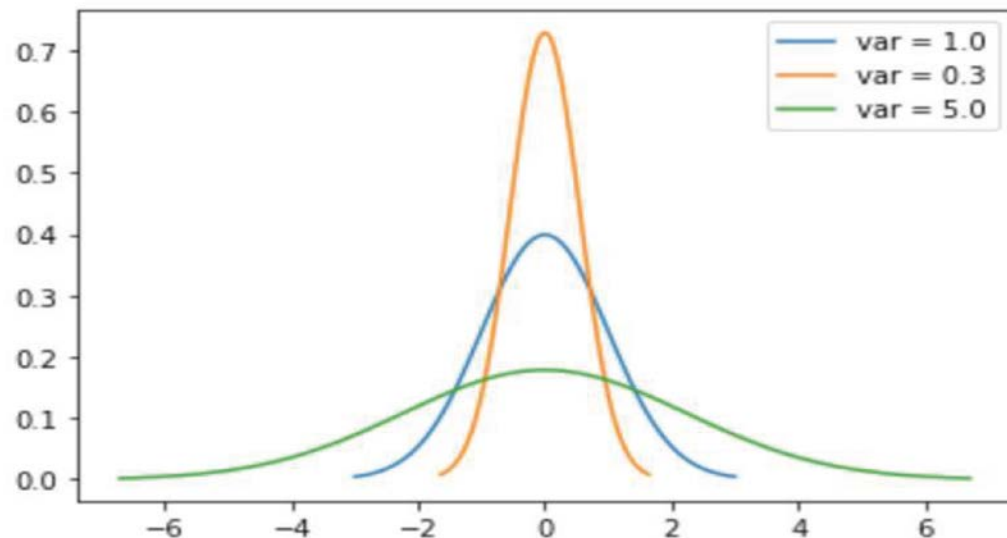
2. 변동성 척도

- **사분위수(quartile):** 자료를 크기 순으로 늘어 놓은 후 똑같은 크기의 네 덩어리로 만들 때 그 경계에 해당하는 값
- **사분위간 범위 (Interquartile range, IQR):**
Q3 ~ Q1의 범위 (범위 = 최대값 - 최소값)
 - 범위가 같더라도 변동성은 다를 수 있다.

5. 데이터 시각화

3. 그래프 종류 - 기술 통계

변동성 척도



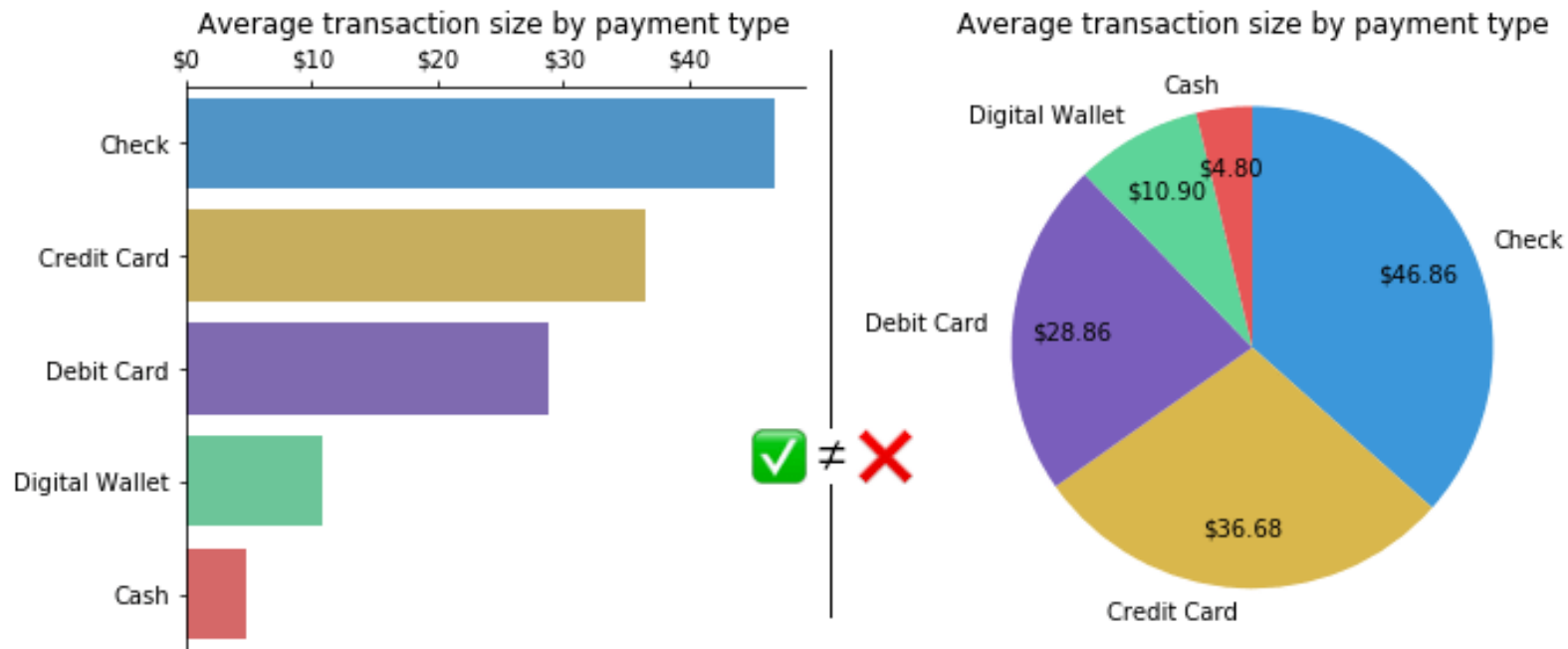
2. 변동성 척도

- **분산(variance):** 자료 각각이 평균으로부터 떨어져 있는 정도
- **표준 편차(standard deviation):** 분산의 제곱근
- **변동 계수(coefficient of variation):** 표준편차를 평균에 대한 상대적인 값으로 표현

5. 데이터 시각화

3. 그래프 종류 - 표현 방식

범주형 자료의 도수 표현 방식 - 바 차트(Bar Chart) & 파이 차트 (Pie Chart)

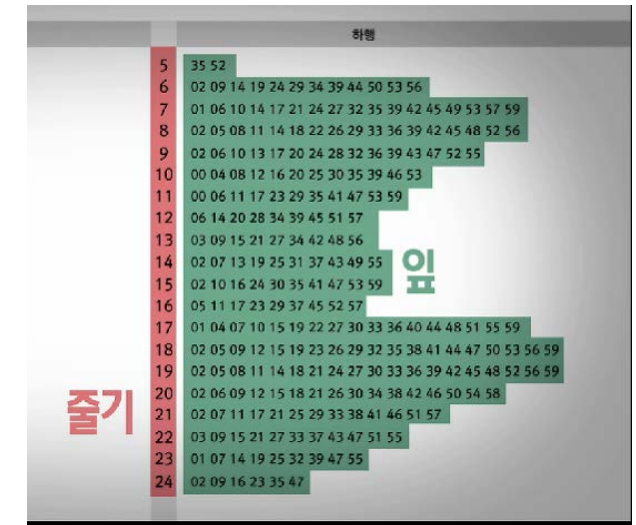
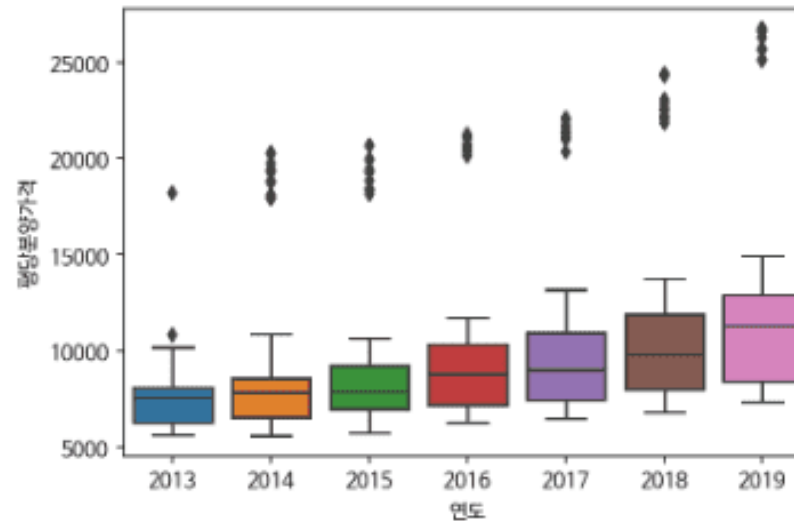
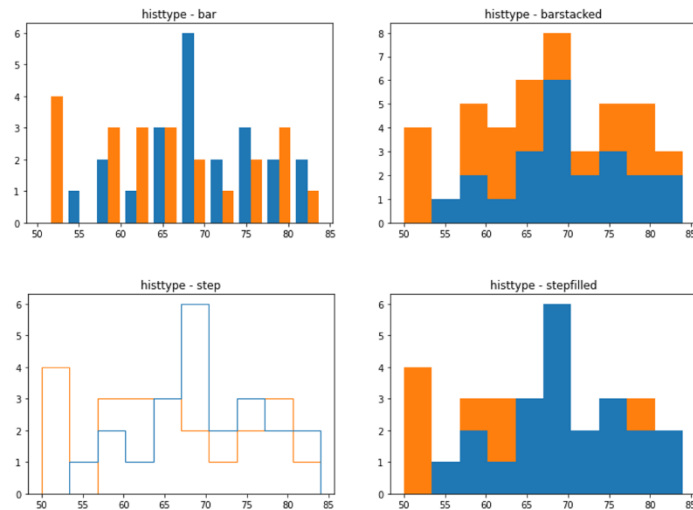


시각화의 목적 중요!: 순위를 파악하고 싶었다면 Bar Chart / 구성 범위를 보고 싶으면 파이차트

5. 데이터 시각화

3. 그래프 종류 - 표현 방식

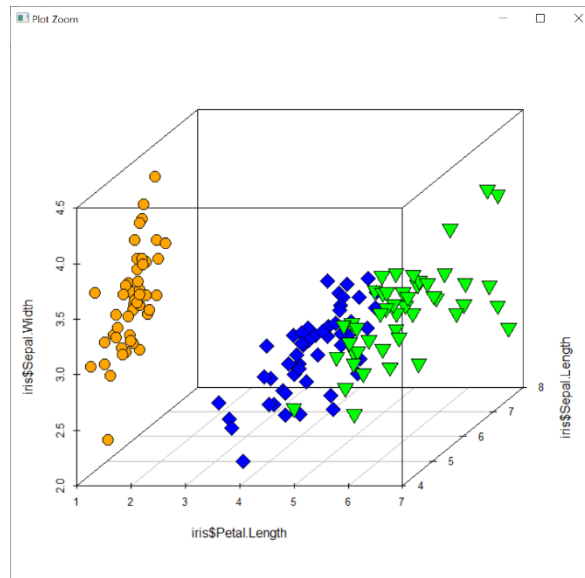
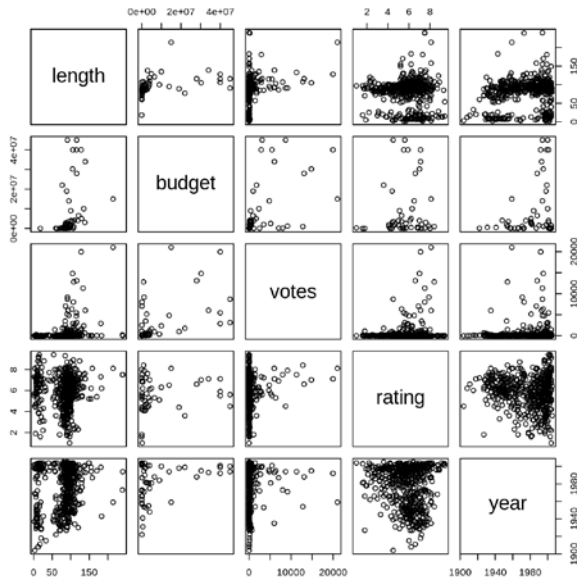
데이터의 분포 확인 - 히스토그램, 박스 plot, 줄기-잎 그림



5. 데이터 시각화

3. 그래프 종류 - 표현 방식

변수간 관계 파악 - 산점도



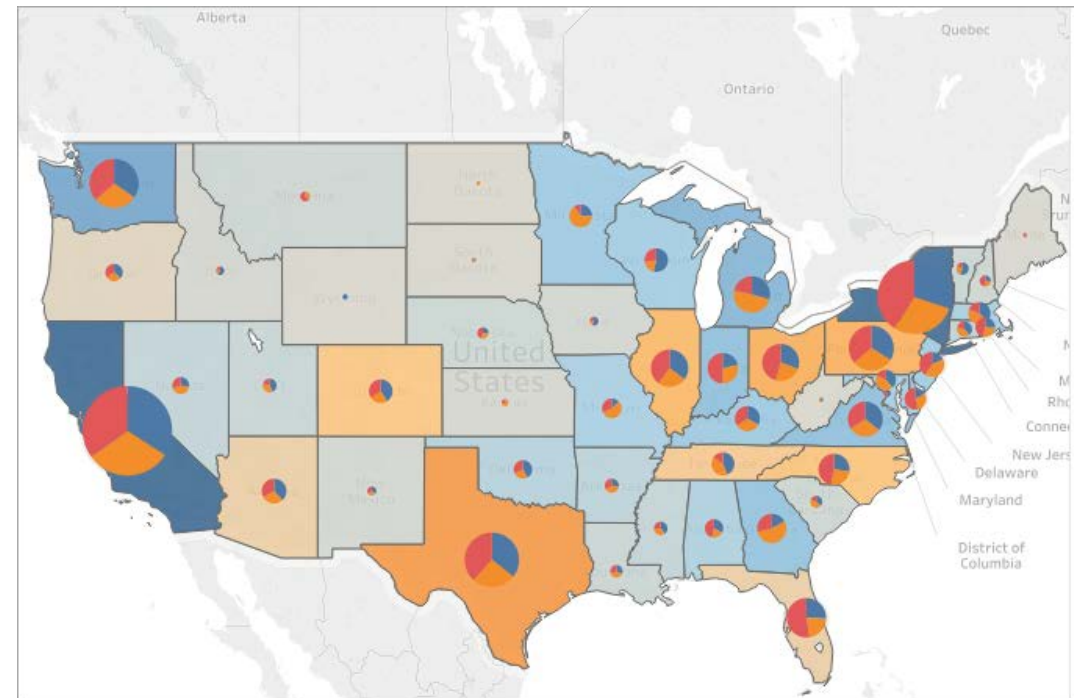
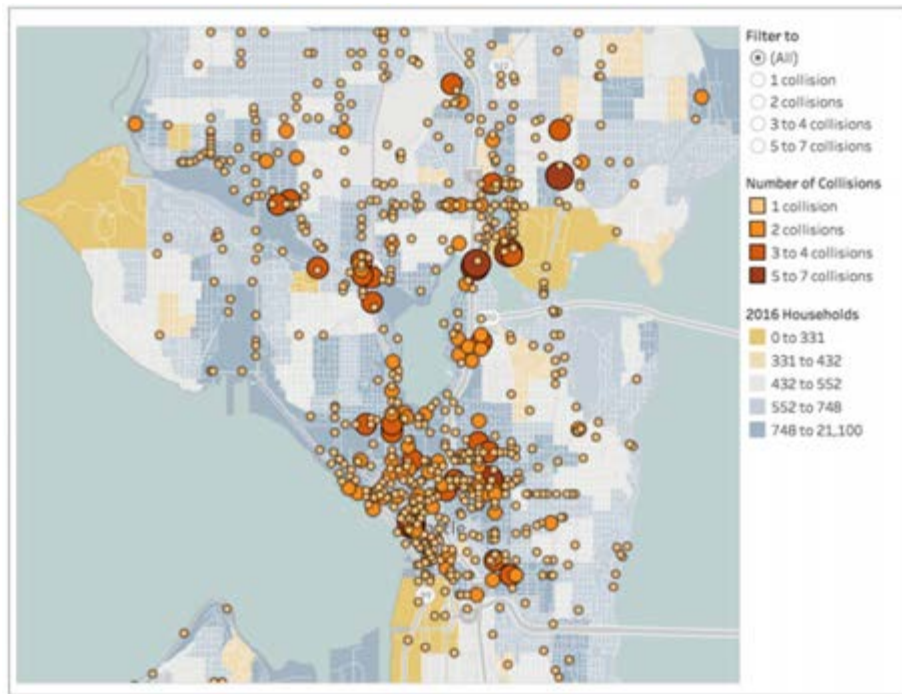
산점도(Scatter Plot)

- 숫자형 변수의 관계를 보기 위해 평면에 관측점을 찍어 만든 통계 그래프
- 변수의 관계와 이상치 유무 등 확인 가능

5. 데이터 시각화

3. 그래프 종류 - 표현 방식

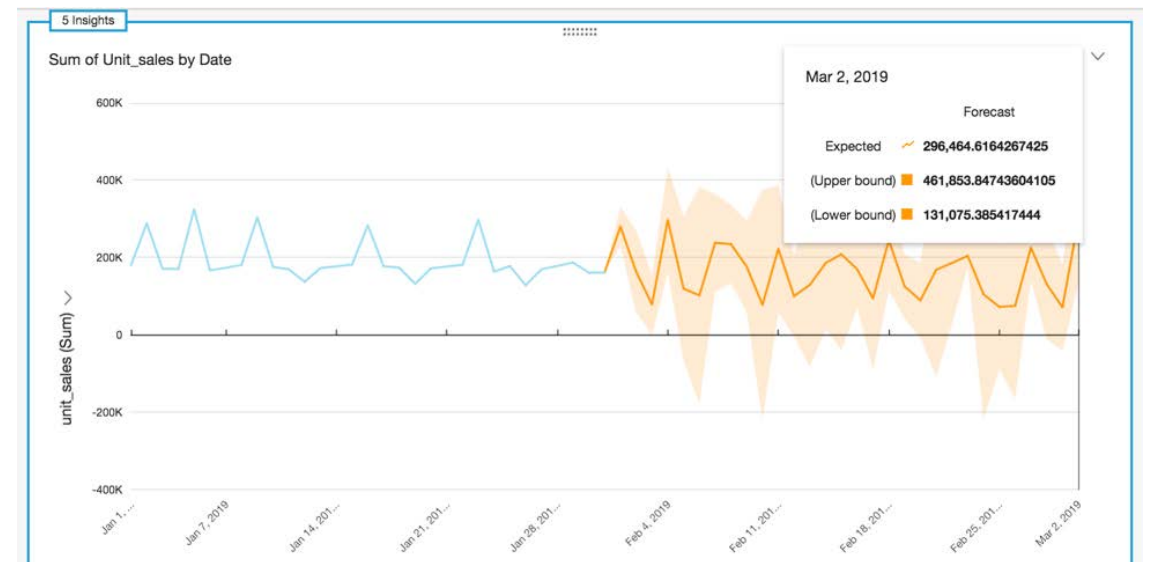
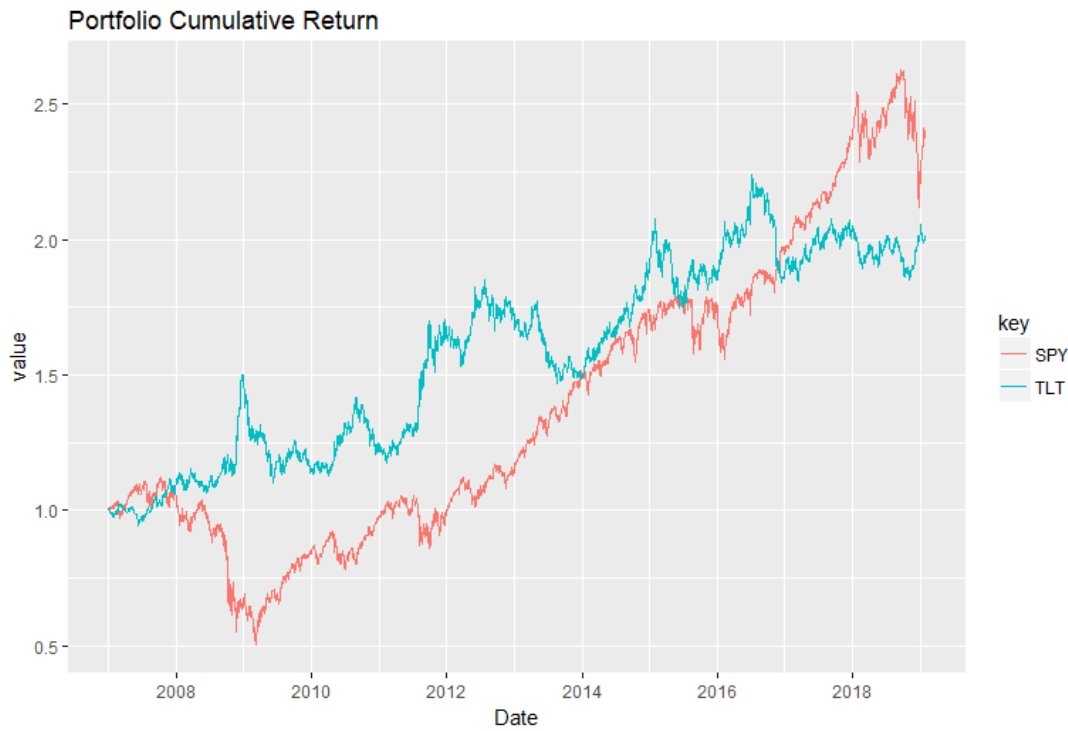
데이터 자료형에 따른 시각화 - Geographic Graph



5. 데이터 시각화

3. 그래프 종류 - 표현 방식

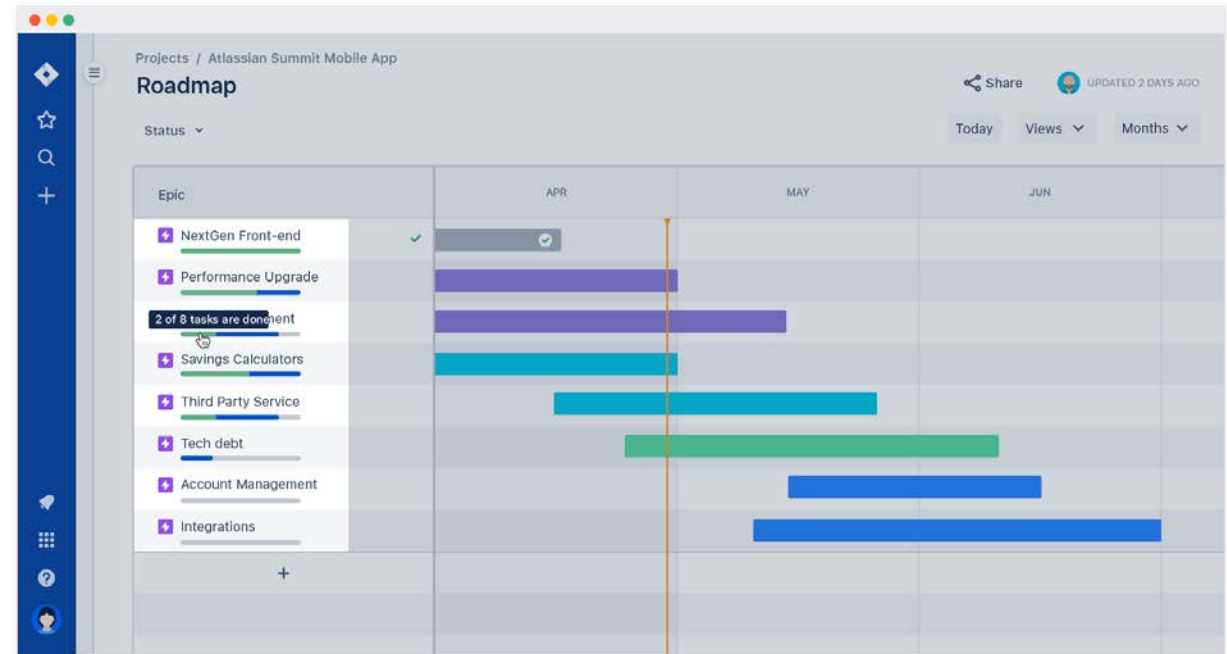
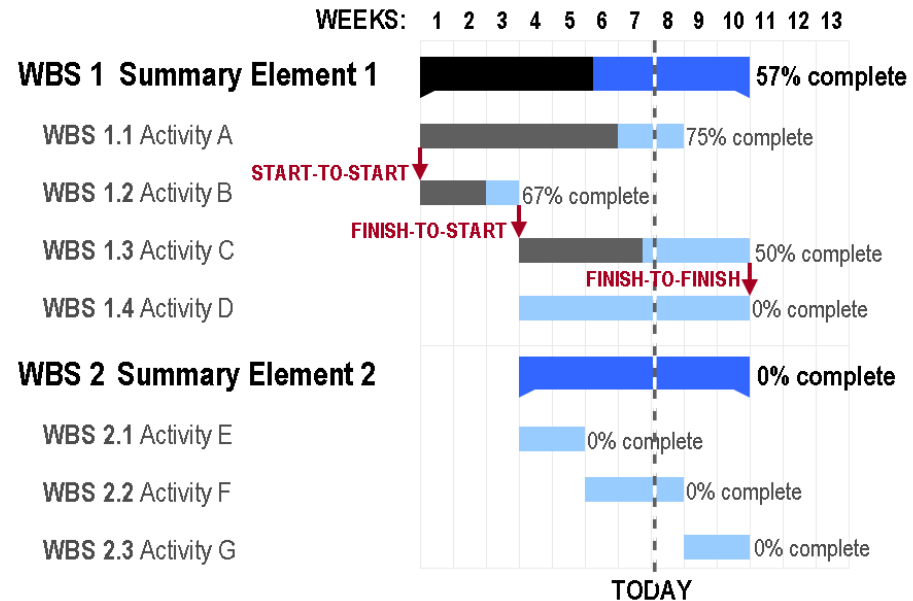
데이터 자료형에 따른 시각화 - 시계열 그래프



5. 데이터 시각화

3. 그래프 종류 - 표현 방식

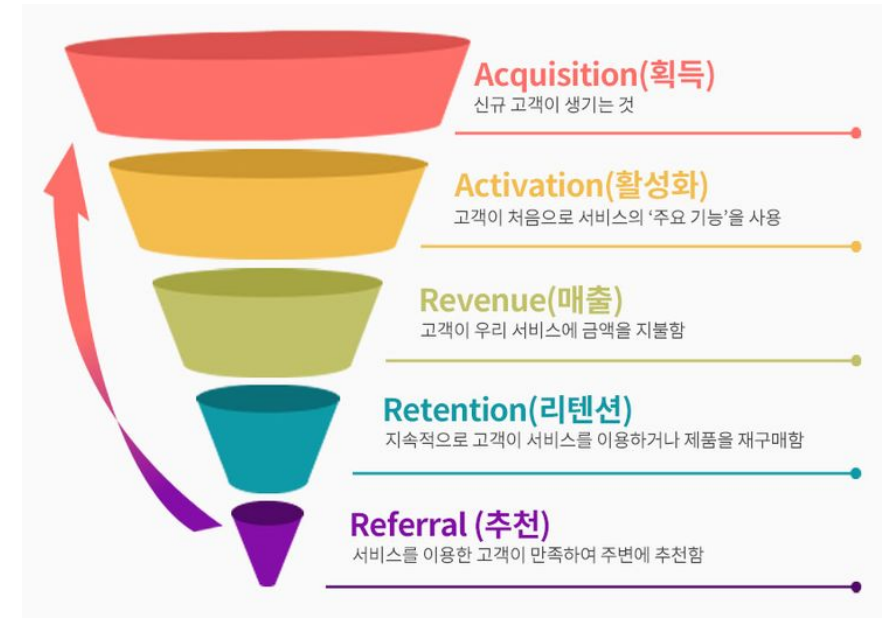
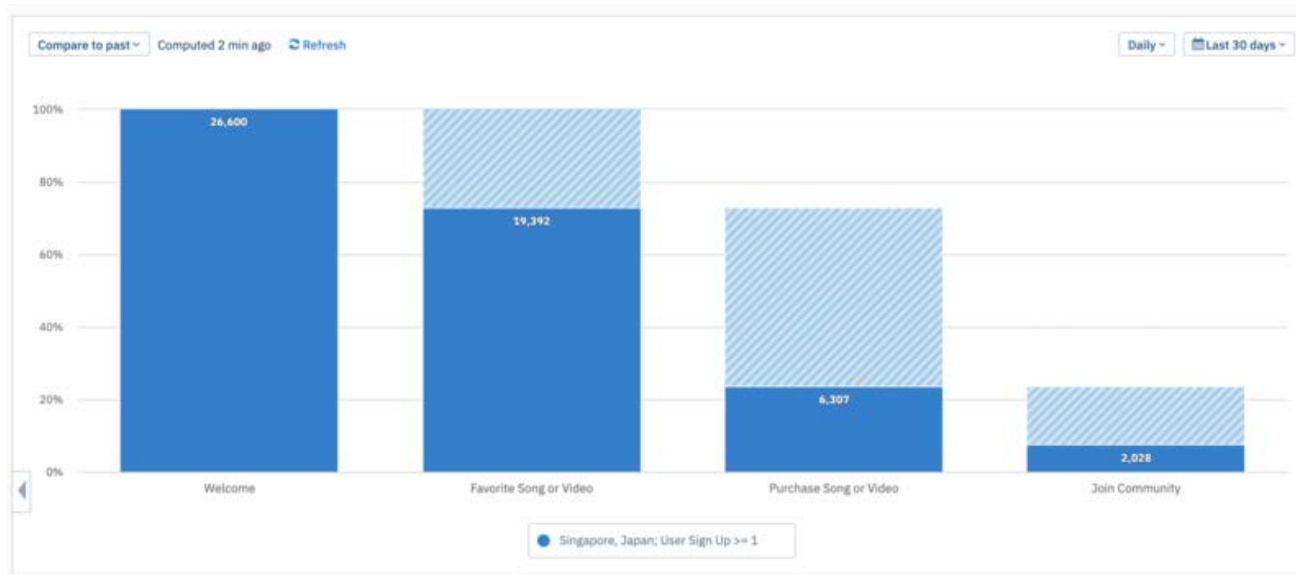
목적에 따른 시각화 - 간트 차트



5. 데이터 시각화

3. 그래프 종류 - 표현 방식

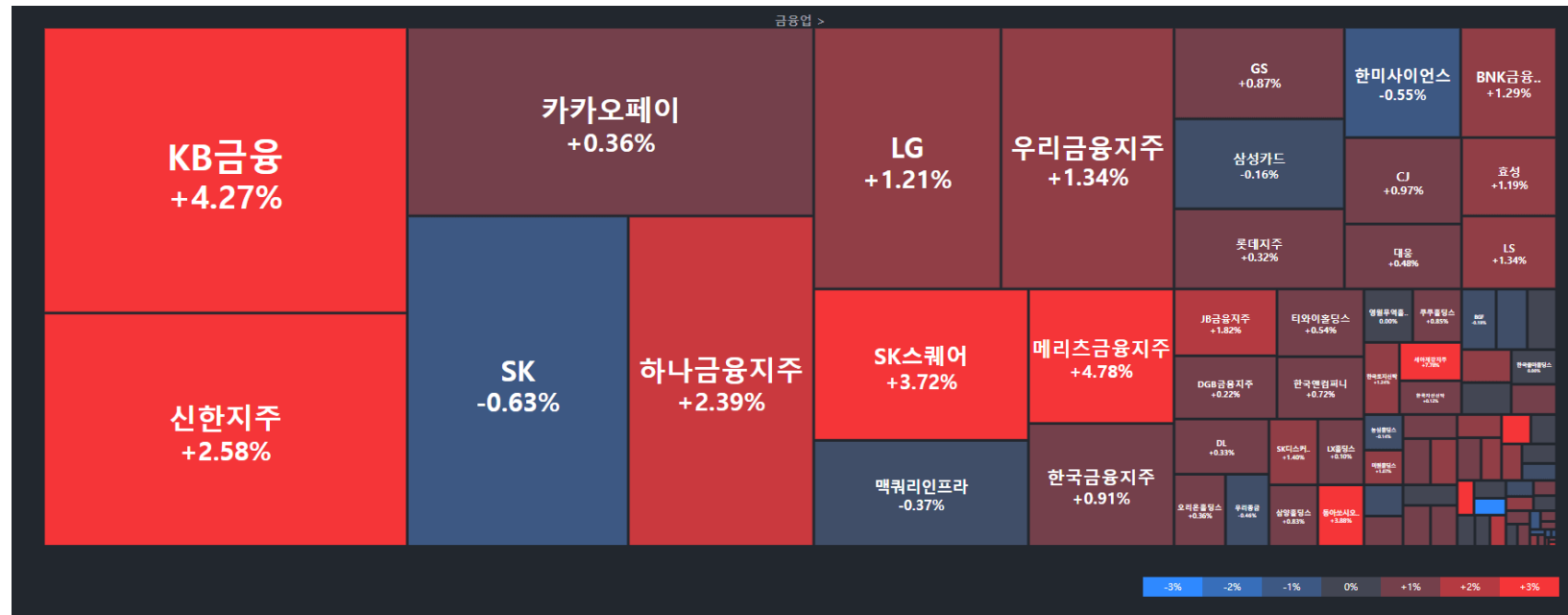
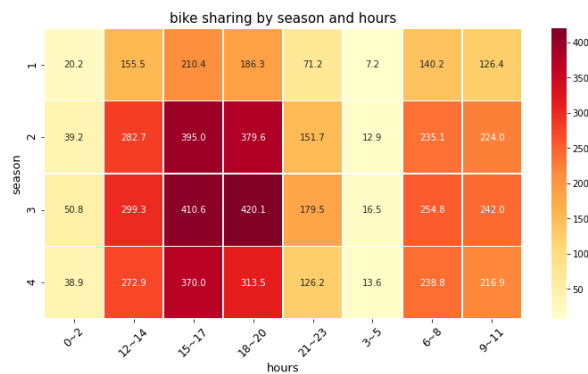
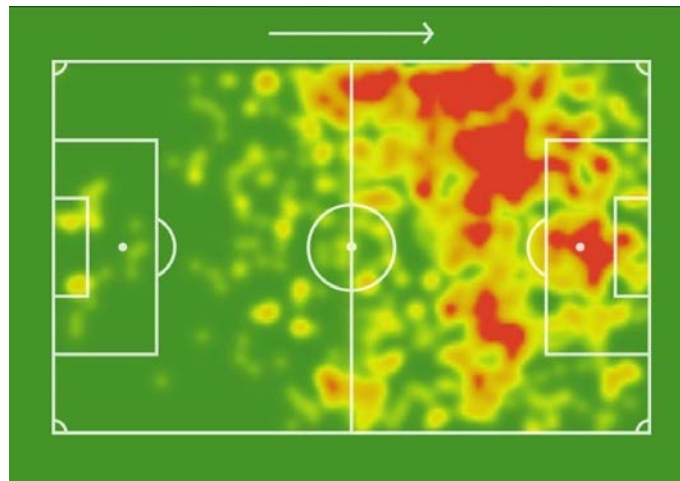
목적에 따른 시각화 - 퍼널 차트



5. 데이터 시각화

3. 그래프 종류 - 표현 방식

기타 시각화 - 히트맵, 트리맵





6. EDA 실습 TIP!

6. EDA 실습 Tip!

YONSEI Data Science Lab | DSL

0. 목적 아이디어이션!

쉬는 시간동안 현재까지 정리된 프로젝트 목표를 수합하고 간단히 소개해주세요!

6. EDA 실습 Tip!

1-1. 데이터별 아이디어이션 - A조

심리 성향 예측 데이터

상세

train

test_x

sample_submission

≡ Views

Grid view

⋮

Hide fields

Filter

Group

Sort

≡

index

QaA

QaE

QbA

QbE

QcA

QcE

1

0

3

363

4

1370

5

997

2

1

5

647

5

1313

3

338

3

2

4

1623

1

1480

1

102

4

3

3

504

3

2311

4

992

5

4

1

927

1

707

5

556

6

5

2

834

1

1769

4

210

7

6

1

1382

1

1473

5

147

8

7

1

384

1

908

5

870

9

8

5

795

2

3469

4

169

10

9

2

1668

1

866

1

895

11

10

1

1465

1

7581

4

134

21 records

Airtable

Copy base

View larger version

- Index
- Q_A / Q_E (a~t) 비식별화를 위해 일부 질문은 Secret 처리
 - Qa : **Secret**
 - Qb : The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught.
 - Qc : Anyone who completely trusts anyone else is asking for trouble.
 - Qd : **Secret**
 - Qe : P.T. Barnum was wrong when he said that there's a sucker born every minute.
 - Qf : There is no excuse for lying to someone else.
 - Qg : **Secret**
 - Qh : Most people forget more easily the death of their parents than the loss of their property.
 - Qi : **Secret**
 - Qj : It is safest to assume that all people have a vicious streak and it will come out when they are given a chance.
 - Qk : All in all, it is better to be humble and honest than to be important and dishonest.
 - Ql : **Secret**
 - Qm : It is hard to get ahead without cutting corners here and there.
 - Qn : **Secret**
 - Qo : The best way to handle people is to tell them what they want to hear.
 - Qp : **Secret**
 - Qq : Most people are basically good and kind.
 - Qr : One should take action only when sure it is morally right.
 - Qs : It is wise to flatter important people.
 - Qt : **Secret**

1=Disagree, 2=Slightly disagree, 3=Neutral, 4=Slightly agree, 5=Agree.

- Q_E(a~t) : 질문을 답할 때까지의 시간

- 비식별 처리가 포함되어 있는 심리 관련 질문 (EX. Qo : 사람을 대하기 가장 쉬운 방법은 듣고 싶은 말을 해주는 것이다.)
- 응답 / 응답까지 소요된 시간
- 5점 척도 (1 ~ 5 : Disagree ~ Agree)

6. EDA 실습 Tip!

1-1. 데이터별 아이디어이션 - A조

심리 성향 예측 데이터

engnat	familysize	gender	hand	married	race	religion
1	4	Female	1	3	White	Other
2	3	Female	1	1	Asian	Hindu
1	3	Male	1	2	White	Other
2	0	Female	1	1	Asian	Hindu
1	2	Male	1	2	White	Agnostic
1	6	Female	1	3	White	Other
1	3	Male	1	1	White	Atheist
1	1	Male	1	1	White	Christian_Other
2	0	Female	2	1	Other	Christian_Other
1	3	Female	1	1	White	Christian_Other
2	2	Female	1	1	Asian	Agnostic
1	2	Male	1	1	White	Christian_Catholic
2	4	Male	1	1	Asian	Muslim
2	2	Female	1	1	Asian	Buddhist
1	3	Male	1	3	White	Christian_Other
1	4	Female	1	2	White	Atheist
1	4	Male	1	2	White	Other

- wr_(01~13) : 실존하는 해당 단어의 정의를 읽
 - 1=Yes, 0=No
- wf_(01~03) : 허구인 단어의 정의를 읽
 - 1=Yes, 0=No

- voted (타겟): 지난 해 국가 선거 투표 여부
 - 1=Yes, 2=No

- tp__(01~07) : items were rated "I see myself as:" _____ such that
 - tp01 : Extraverted, enthusiastic.
 - tp02 : Critical, quarrelsome.
 - tp03 : Dependable, self-disciplined.
 - tp04 : Anxious, easily upset.
 - tp05 : Open to new experiences, complex.
 - tp06 : Reserved, quiet.
 - tp07 : Sympathetic, warm.
 - tp08 : Disorganized, careless.
 - tp09 : Calm, emotionally stable.
 - tp10 : Conventional, uncreative.

데이터 설명

- 응답자의 Demographic별 특징 포함
- 응답자의 답변 신뢰도 관련 문항

데이터 목적

- 해당 설문자의 국가 선거 투표 여부 알고리즘 작성

6. EDA 실습 Tip!

1-1. 데이터별 아이디어이션 - A조

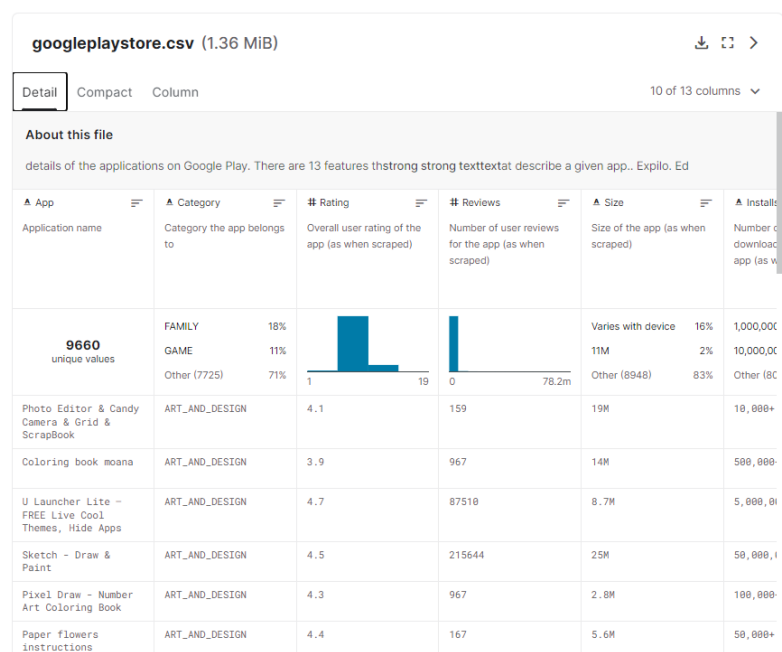
심리 성향 예측 데이터

- 예시 목표1 : 선거 투표 여부를 맞추는 알고리즘을 만들었을 때 어떤 요소들이 중요하게 작용했을까?
- 예시 목표2 : 언어에 따라 같은 Demographic군의 사람들의 심리 상태가 다를까?
- 예시 방식1 : 변수간 관계 확인 시각화 (X to Y) / 통계적 접근: 요인 분석 (ex. 1순위: age_group, 2순위: Qb, 3순위 ...)
- 예시 방식2 : 데모그래픽 집단을 나눌 상관관계 확인 -> 라벨링 -> 심리 상태 분포 확인 -> 언어별 차이 확인

6. EDA 실습 Tip!

1-2. 데이터별 아이디어이션 - B조

Google playstore 데이터



6. EDA 실습 Tip!

1-2. 데이터별 아이디어이션 - B조

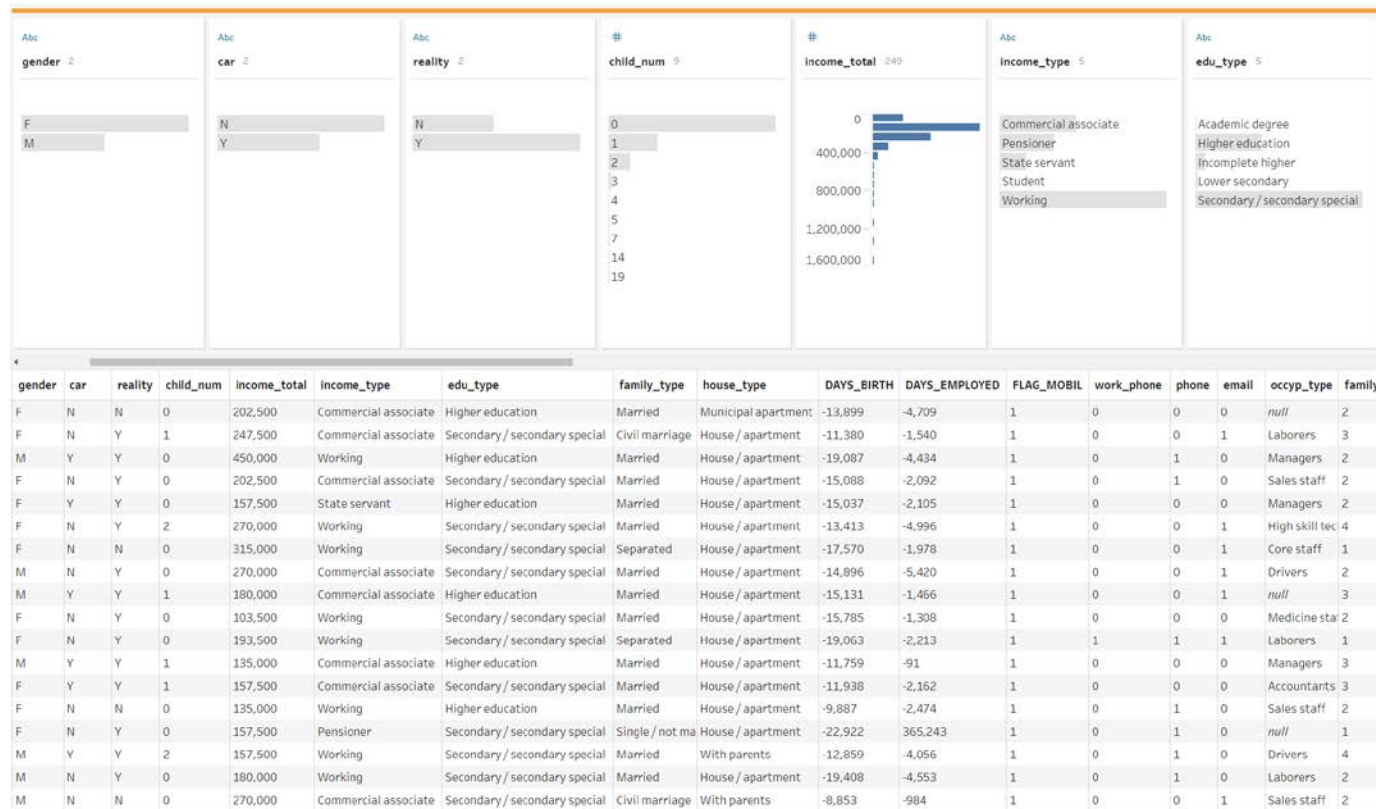
Google playstore 데이터

- 예시 목표1 : 앱의 카테고리별로 **평점 분포가 다르게** 나타날까? 만약 그렇다면 그 이유는 무엇일까?
- 예시 목표2 : 구글 스토어 평점 분포와 **앱스토어(애플)의 평점 분포가 동일할까?**
- 예시 방식1 : 평점 분포가 가장 유사하게 나타나게 하는 변수와 아닌 **변수 확인** / 요인, 상관 분석
- 예시 방식2 : **앱스토어 데이터 크롤링** 후 분석 결과 비교
- P.S. 리뷰 감정 양극화 정도 & 다운로드 등의 변수들을 통해서 리뷰를 구매했을 것 같은 어플 예측 (- 모델링)

6. EDA 실습 Tip!

1-3. 데이터별 아이디어이션 - C조

신용카드 연체 예측 데이터



데이터 설명

- Gender : 성별
- car : 차량 소유 여부
- reality : 부동산 소유 여부
- child_num : 자녀 수
- income_total : 연간 소득
- income_type : 소득 분류
- edu_type : 교육 수준
- family_type : 결혼 여부
- house_type : 생활 방식
- DAYS_BIRTH : 출생일
- DAYS_EMPLOYED : 업무 시작일
- FLAG_MOBIL : 핸드폰 소유 여부
- work_phone : 업무용 전화 소유 여부
- phone : 전화 소유 여부
- email : 이메일 소유 여부
- occyp_type : 직업 유형
- Family_size : 가족 규모
- begin_month : 신용카드 발급 월
- credit : 사용자의 신용카드 대금 연체를 기준으로 한 신용도

목적: 신용카드 사용자의 개인정보와 데이터를 이용하여 향후 채무 불이행과 대금 연체 가능성을 예측

6. EDA 실습 Tip!

1-3. 데이터별 아이디어이션 - C조

신용카드 연체 예측 데이터

- 예시 목표1 : 연체 가능성을 높이는 변수는 무엇일까? 변수간 영향 정도를 순위화할 수 있을까?
- 예시 목표2 : 교육 수준에 따라 결혼여부가 달라질까? / Stereo Type을 깨는 다른 결과는 무엇이 있을까?
- 예시 방식1 : 변수간 관계 확인 시각화 (X to Y) / 통계적 접근: 요인 분석 (ex. 1순위: age_group, 2순위: Qb, 3순위 ...)
- 예시 방식2 : CDA에 가깝지만 질문을 던지면서 이후 분석을 진행할 여지를 확인

6. EDA 실습 Tip!

1-4. 데이터별 아이디어이션 - D조

Youtube 데이터


KRvideos.csv (34.84 MiB)

Detail

Compact

Column

10 of 16 columns

video_id	trending_date	title	channel_title	category_id	publist	
#NAME?	1%	205 unique values	16353 unique values	4043 unique values		29Sep11
AKDuAzSwaPI	0%					
Other (34114)	99%					
RxGQe4EeEpA	17.14.11	좋아 by 민서_윤종신_듣니 답가	리무이코리아	22	2017-11-13T07:08	
hh7wVE801Q8	17.14.11	JSA 귀순 북한군 총격 두 상	Edward	25	2017-11-13T10:56	
9V8bnWUmE9U	17.14.11	나물라패밀리 운동화 영상 2단 (빠빠로데이버전)	나물라패밀리 핫쇼	22	2017-11-11T07:16	
0_8py-t5R88	17.14.11	이명박 출국 현장, 놓치면 안되는 장면	미디어통구	25	2017-11-12T11:16	
bk55Rbx1QdI	17.14.11	갑장검은 물려갔다 MBC 노 조 환호와 눈물	NocutV	25	2017-11-13T11:08	
AmP0ryzDmbY	17.14.11	김정숙 여사는 왜 갑자기 문재인 대통령 주머니에 손을 넣었나? 인도네시아 대통령도 깜놀	하우스	25	2017-11-12T10:17	
4Nxb_nQDYWo	17.14.11	예능신 이광수 하이라이트 모음	채란이의 즐거운 유튜브 모음	22	2017-11-12T03:36	

Data Explorer

Version 115 (539.22 MiB)

{}

 CA_category_id.json

CAvideos.csv

{}

 DE_category_id.json

DEvideos.csv

{}

 FR_category_id.json

FRvideos.csv

{}

 GB_category_id.json

GBvideos.csv

{}

 IN_category_id.json

INvideos.csv

{}

 JP_category_id.json

JPvideos.csv

{}

 KR_category_id.json

KRvideos.csv

{}

 MX_category_id.json

MXvideos.csv

{}

 RU_category_id.json

RUvideos.csv

{}

 US_category_id.json

USvideos.csv

Summary

20 files

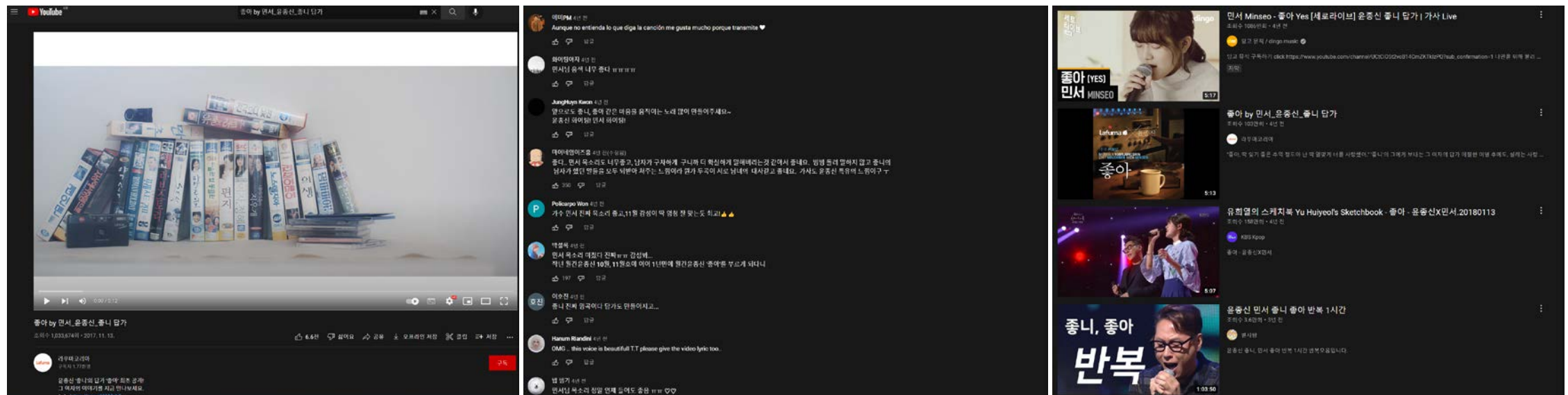
데이터 설명

- Field가 국가마다 있음
- 이 중에서 한국 데이터 사용
- 비디오 명 / 채널 / 트렌딩 일자 / 카테고리 / 게시
일자 / 시청 수 / 좋아요 / 싫어요

6. EDA 실습 Tip!

1-4. 데이터별 아이디어이션 - D조

Youtube 데이터



- 실제 페이지를 찾아갈 수 있음
- 런타임 데이터(시간/분/초), 댓글 수 / 내용 / 좋아요 최다 댓글 (감정) / 좋아요 N개 이상 데이터 개수
- 같은 검색어 유사 동영상 여부/개수/조회수 차이
- 영상 설명 데이터

6. EDA 실습 Tip!

1-4. 데이터별 아이디어이션 - D조

Youtube 데이터

- 예시 목표1 : 카테고리별로 트렌딩하는 제목은 무엇일까?
- 예시 목표2 : 같은 주제 / 유사 카테고리의 동영상의 순위를 가르는 다른 요소는 무엇일까?
- 예시 방식1 : 제목을 토큰화하여 카테고리별로 유의미하게 다른 트렌딩 키워드 확인
- 예시 방식2 : 크롤링을 통해 영상 길이, 종댓구알 등의 데이터를 가져와서 비교
- PS. 논란 영상이 트렌딩할 때, 사회적 이슈의 트렌딩과 관련이 있을까? (뉴스 제목 크롤링 - 정치&연예 etc..)

6. EDA 실습 Tip!

1-5. 데이터별 아이디어이션 - E조

소상공인시장진흥공단 상권 데이터

#	상가업소번호 10만	상호명 8만	지점명 6만	상권업종대분류코드 8	상권업종중분류명 8	상권업종중분류코드 89	상권업종중분류명 89
상가업소번호	상호명	지점명	상권업종대분류	상권업종중분류	상권업종중분류명	상권업종중분류코드	상권업종중분류명
25,033,300	동그라미중고타이어	null	D	소매	D23	자동차/자동차	D23A04
17,174,549	세인트존스호텔Ohcrab	null	O	숙박	O01	호텔/콘도	O01A01
17,174,079	평정리마디호텔	null	O	숙박	O01	호텔/콘도	O01A01
17,173,904	호텔합스텐스카이라운지	null	O	숙박	O01	호텔/콘도	O01A01
24,412,526	레이디가구	null	D	소매	D15	가구소매	D15A01
24,611,751	엘로디피아노교습소	null	R	학문/교육	R09	학원(종합)	R09A01
17,174,157	지진플러스원주기업도시점	원주기업	Q	음식	Q09	유흥주점	Q09A01
17,174,566	307포차	null	Q	음식	Q09	유흥주점	Q09A07
21,651,344	영진보일러	null	D	소매	D21	철물/난방/냉	D21A02
21,651,627	금강산시유리공사	null	D	소매	D21	철물/난방/냉	D21A08
24,652,566	유성가스	null	F	생활서비스	F16	주유소/충전	F16A02
25,114,427	스카이발상회	null	D	소매	D06	가방/신발/의	D06A07
20,782,013	영복나눔	null	D	소매	D07	가정/주방/의	D07A05
20,778,927	커즈카페야사탕	null	Q	음식	Q12	커피점/카페	Q12A01
17,175,358	국수나루	null	Q	음식	Q01	한식	Q01A01
17,175,250	꼭지네식당	null	Q	음식	Q01	한식	Q01A01
17,175,299	죽발야시장	null	Q	음식	Q01	한식	Q01A08
21,691,865	LEADUS	null	D	소매	D23	자동차/자동차	D23A07

데이터 설명

- 상가업소번호
- 상호명
- 지점명
- 상권업종대분류코드
- 상권업종중분류명
- 상권업종중분류코드
- 상권업종중분류명
- 상권업종소분류코드
- 상권업종소분류명
- 표준산업분류코드
- 표준산업분류명
- 시도코드
- 시도명
- 시군구코드
- 시군구명
- Etc..

6. EDA 실습 Tip!

1-5. 데이터별 아이디어이션 - E조

소상공인시장진흥공단 상권 데이터

CSV

소상공인시장진흥공단_상가(상권)정보

다운로드

요청신고 및 담당자 문의

파일데이터 정보

메타데이터 다운로드

파일데이터명	소상공인시장진흥공단_상가(상권)정보_20220331		
분류체계	산업·통상·중소기업 - 산업·중소기업일반	제공기관	소상공인시장진흥공단
관리부서명	상권분석실	관리부서 전화번호	042-363-7852
보유근거	소기업 및 소상공인 지원을 위한 특별조치법 제13조	수집방법	
업데이트 주기	분기	차기 등록 예정일	2022-07-29
매체유형	텍스트	전체 행	1
확장자	CSV	다운로드(바로가기)	75524
데이터 한계		키워드	상가업소 ,소상공인 ,상권정보
등록	2022-04-26	수정	2022-06-27
제공형태	공공데이터포털에서 다운로드(원문파일등록)		
설명	영업 중인 전국 상가업소 데이터를 제공합니다. (상호명, 업종코드, 업종명, 지번주소, 도로명주소, 경도, 위도 등)		
기타 유의사항	utf-8로 인코딩 되어 utf-8파일 열람방법을 zip파일안에 포함		
비용부과유무	무료	비용부과기준 및 단위	건
이용허락범위	이용허락범위 제한 없음		

데이터 목적

- 데이터 제공 주체가 국가
- 분석 목적의 데이터 아님 (운영 목적의 데이터)
- 어떤 운영 목적이었을까? 고려

운영 목적 데이터 특징

- 정확함
- 연결성이 높다 (해당 조직 내 다른 데이터와 결합 용이)
- 국가 단위 해결 필요 문제 고려 (Specific < Impact)

6. EDA 실습 Tip!

1-5. 데이터별 아이디어이션 - E조

소상공인시장진흥공단 상권 데이터

- 예시 목표1 : 지역별 업종 분포는 어떻게 다를까? / 업종별 최다 업체의 분포가 지역마다 같을까?
- 예시 목표2 : 지역별로 동일 업종의 매출에 차이가 있을까?
- 예시 방식1 : 상호명 전처리 & 업종별 구분한 후 시각화 (위도, 경도 사용하면 지도 그래프도 가능)
- 예시 방식2 : 외부데이터를 확보하여 결합하여 데이터 분석 (크롤링 X)
- PS. 분석 필요 사항을 파악하기 위해 '소상공인' 관련 뉴스 제목 크롤링 후 분석하여 주제를 정하는 것도 방법

(공정 분석: 목표 설정을 위한 분석)

6. EDA 실습 Tip!

2. 프로젝트 진행방식

1. 데이터 셋 설정 (<https://www.notion.so/099b19dac7394b9e9d01e09698a62a68>)
2. 데이터 확인 (확인용 시각화)
3. 목표 설정 & 분석 방향 아이디어이션 (<https://www.notion.so/1-dcc27af434a74592ae6459503e868966>)
4. 데이터 전처리 작업
5. 데이터 시각화 (분석 결과 전달용 시각화)
6. 스토리 라인 작성 후 추가 분석
7. 참고 자료 정리 & 내용 추가 (Action Plan 제안을 위한 도메인 지식 결합)
8. 결과 발표

YONSEI Data Science Lab | DSL

💡 프로젝트 인사이트 제안 방향

- ### 1. (데이터셋 1) 설문조사 데이터를 통한 경쟁사 어플 사용자들의 니즈 확인

- 경쟁사 앱 사용 현황
- 수면 어려움 정도별 앱 사용 현황 / 앱 사용 이유
- 앱 사용 이유별 수면 어려움 정도
- 주 이용 연령층 / 주 이용 성별
- 신분별 수면 어려움 정도 / 이유 / 앱 사용 이유
- 신분별 경쟁사 앱 사용 현황
- 경쟁사 앱 성별/연령/신분 분포

2. (데이터셋 3) AARRR 중 Aquisition 최적화 [마케팅-유입]

3. (리뷰 크롤링 데이터 셋) SWOT 중 OT 집중 ← 경쟁사 대비 비교 우위 선정 가능 지점 확인

- 경쟁사별 리뷰 평점 분포
- 고점 리뷰 이유 NLP
- 저점 리뷰 NLP

4. (리뷰 크롤링 데이터 셋) STP 구체화

P.S. BUT 최고 중요 사항

- 1) 이 회사의 BM이 무엇일까?
- 2) BM을 지향해나갈 수 있는 전략 방향으로 분석을 하면 좋을 것으로 보임.
- 3) BM이 명확하지 않다면 경쟁사 분석을 통해 실현 가능한 BM을 제안해야 함.

→ 어플리케이션의 BM은 크게

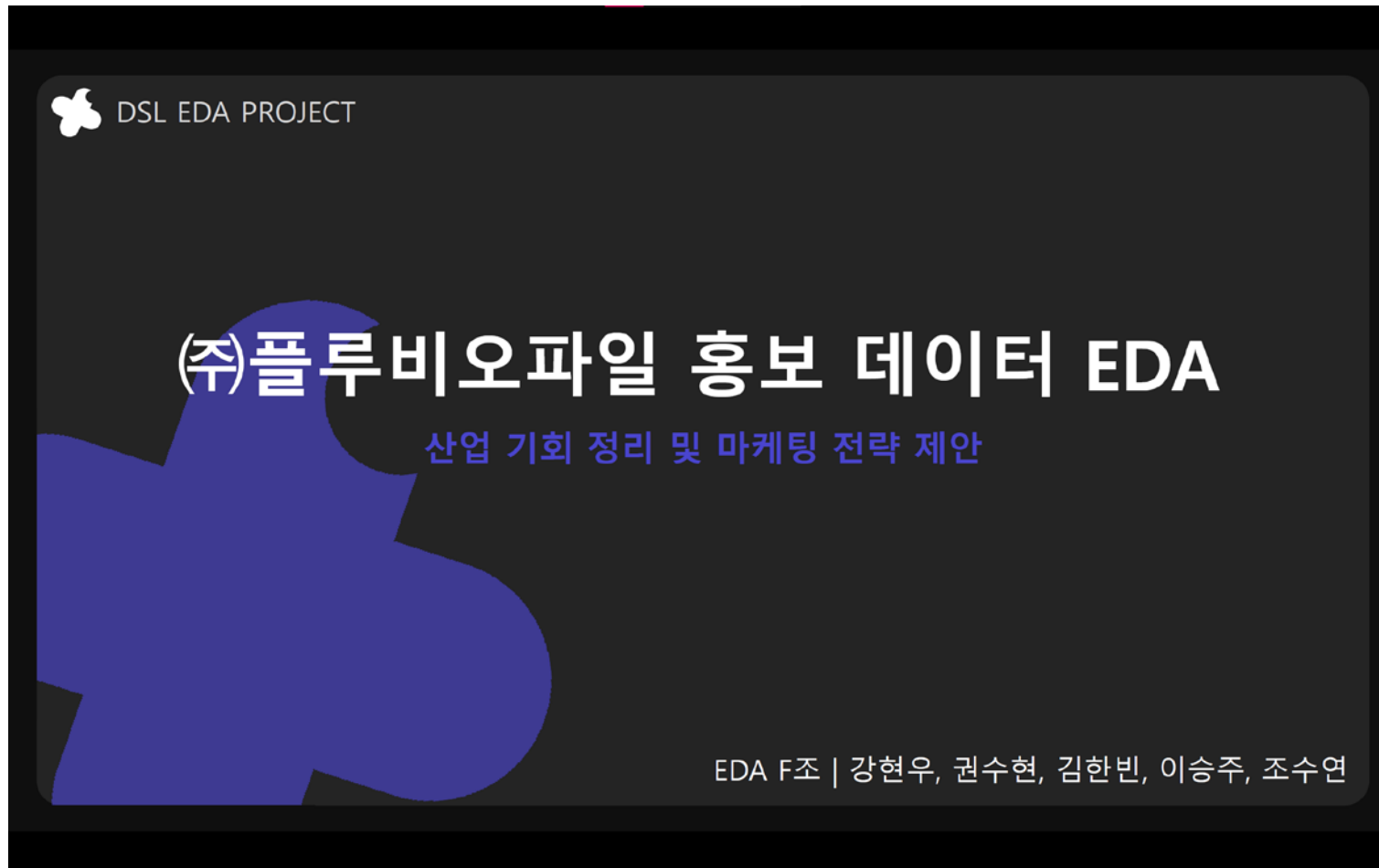
1. 플랫폼(광고업) ← 커뮤니티 형성 가능 여부 파악 필요
2. 플랫폼(상품판매) ← 판매 가능 상품 제작 (가능 여부)
3. 플랫폼(수수료) ← 프리미엄 서비스 제공

61

6. EDA 실습 Tip!

YONSEI Data Science Lab | DSL

3. 프로젝트 발표 예시



6. 세션 과제!

3. 프로젝트 발표 예시

본인 팀의 데이터를 보고 자신이 생각하는 프로젝트 분석 개요를 A4용지 한 장 분량으로 적어서 노선에 업로드해주세요!



감사합니다