The First Data Analysis

올바른 시작점 & 분석 한 바퀴

D:plus

목차

D:plus

1. 현실로부터

1. 현실로 부터.

학술적인 분류가 아니라, 데이터 분석절차에 맞추어 통계적인 방법들을 바라볼 필요가 있다.

1. [의문 / 가설 / 아이디어 / 주제] 가 있는가? 없는가?
2. 의문의 네 가지 타입
3. 네 가지 타입에 따른 방법론

그리고, 계속 머리 속에 가지고 있어야 할 생각이 있다.

Signal & **Noise** :

What is the **Signal** and What is the **Noise**?

D:plus

1. 현실로 부터.

학술적인 분류가 아니라, 데이터 분석절차에 맞추어 통계적인 방법들을 바라볼 필요가 있다.

1. [의문 / 가설 / 아이디어 / 주제] 가 있는가? 없는가?
2. 의문의 네 가지 타입
3. 네 가지 타입에 따른 방법론

그리고, 계속 머리 속에 가지고 있어야 할 생각이 있다.

Signal & **Noise** :

What is the **Signal** and What is the **Noise**?

D:plus

1. 현실로 부터.

의문 / 가설 / 아이디어 / 주제 가 있는가? 없는가?

Type A

- **데이터만 있다.**
- 뭘 분석해야할지 모르겠다.
- 뭔가 하고 싶은데, 뭐 하고 싶지?

Type B

- **추상적이지만 가설/의문이 있다.**
- 목표가 있다.
- 데이터도 있다.

1. 현실로 부터.

두 타입 모두 결국은

ㄱ. 목표 재설정 : 해결 가능한 수준부터 단계적으로.

ex> 전체 매출 10% 성장 -> 작년 대비 ROI 10% 성장 -> …

ㄴ. 단계별로 가설 설정 -> 탐색 -> 가설 설정 -> 탐색 -> …

ex> 회원 가입율 문제를 해결해야 한다.

1. 타고 들어오는 배너에 문제가 있을까? -> 문제 있긴 있는 듯 -> 배너의 디자인이 문제인가?
2. 회원 가입 경로가 너무 긴가?
3. 디바이스에 따른 차이가 있는 걸까? -> 디바이스에 따라 회원 가입 차이가 있는 듯
   -> A디바이스는 뭐가 문제길래 회원 가입율이 저조한 거야? ->
4. UI가 너무 사람들의 주의를 분산시키는가? -> UI를 바꾸면 뭔가 개선 될까?

확인해봐야 하는 구체적인 지점을 우선순위대로 나열할 수 있을 때 까지.

ㄷ. 실험계획 / 모델링

아래와 같은 수준으로 정리가 되면 진행할 수 있다.

ex> UI개선안1, 개선안2, 기존안 셋 중 어떤 것이 가장 회원 가입율이 높은가?

ex> 이 상품의 다음주 판매량은 어떻게 되는가?

ㄹ. 해당 단계 목표가 해결될 때까지 ㄴ, ㄷ 반복

D:plus

1. 현실로 부터.

거칠게 간추려보자.

1. 목표에 맞추어 어떻게 의문/가설/질문 을 만들어내고 그것을 점점 구체화 시키는가.

2. 액션에 돌입하기 전에, 실험을 계획할 가설이 있는가.

3. 액션으로 이어질, 모델이 있는가.

1. 현실로 부터.

더 거칠게 간추려보자.

의문 / 가설 / 질문을
어떻게 만들고 어떻게 확인하는가.

이 강의에서 고민하게 되는 것!

자연스럽게 떠오르는, 영감을 받아 떠오르는 의문 / 가설들을 더욱 구체화 하는 방법

의문 / 가설들이 자연스럽게 떠오르게 만들 관점.

D:plus

1. 현실로 부터.

이 강의에서 첫 번째로 고민하게 되는 것!

자연스럽게 떠오르는 의문 / 가설들을 구체화 하는 방법
의문 / 가설들이 자연스럽게 떠오르게 만들 관점.

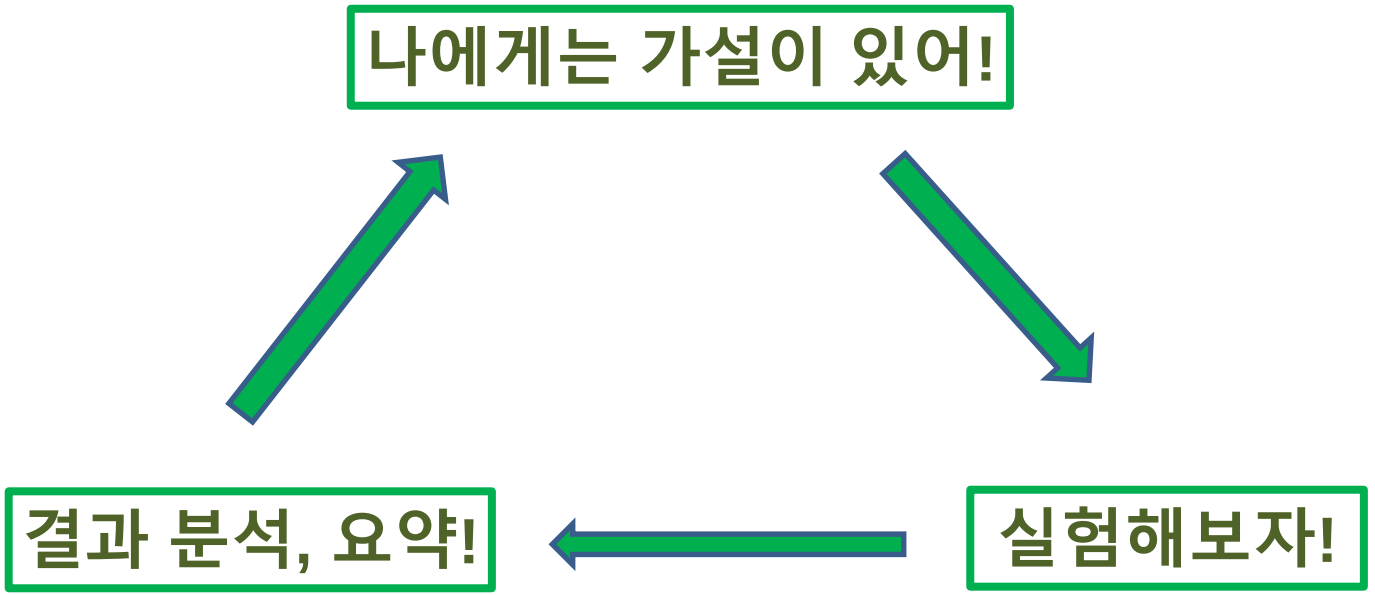
D:plus

2. Exploratory Data Analysis

- ✓ Why EDA?
- ✓ 4 types of Questions

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석



2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

나에게는 가설이 있어!

Confirmatory Data Analysis
통계의 중심이 될 수 밖에 없었다.

결과 분석, 요약! ← 실험해보자!

Note : Confirmatory Data Analysis, 확증적 데이터 분석.

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

가설 그런 것은 넘쳐남.

실험을 통해 가설을 확인하는 것이 문제

통계는 일상

SCIENTISTS

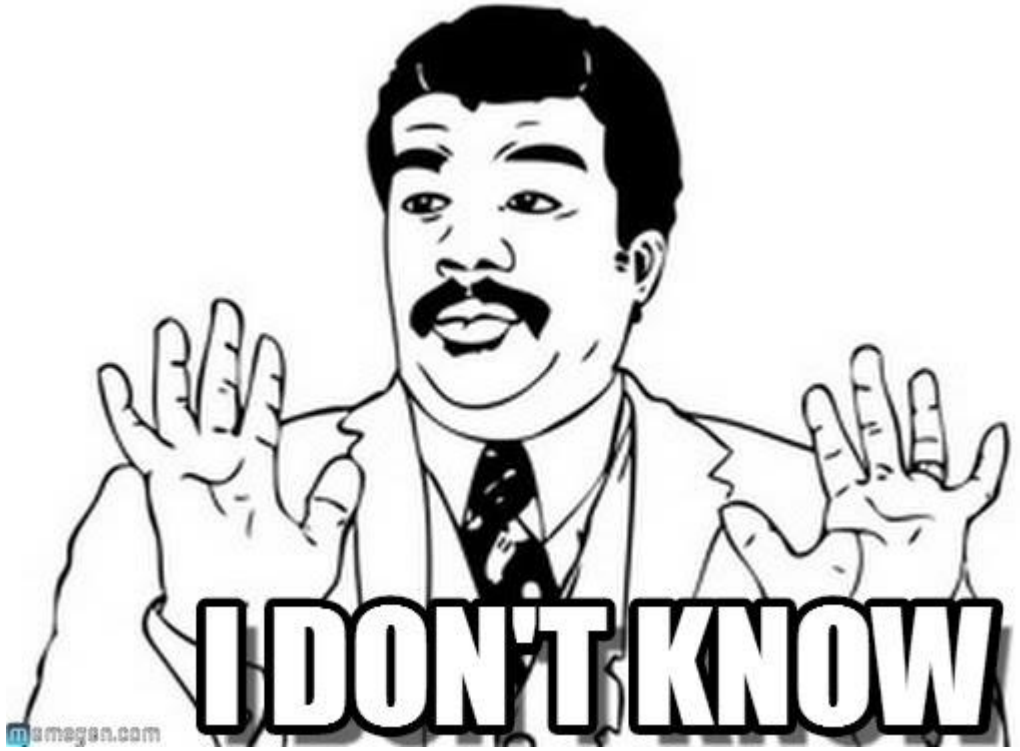
~~Note : 사실 자기가 사용하는 도구가 통계인지 잘 인식 못함.~~
~~확증적 데이터 분석 그런 말도 잘 모름.~~
~~P-value구한다고 하면 귀신 같이 알아들음.~~
~~근데, P-value가 뭔지는 또 잘 모름.~~

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

하지만 보통은 뭘 분석 해야 할지 **모름**.
어디서부터 시작해야 할지 **모름**.
사실 데이터 어디 있는지도 **모름**.



가설 그런 거 없음

뭘 어떻게?

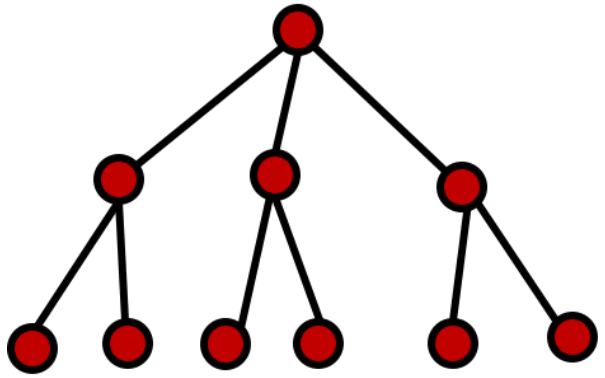
분석해보자!

D:plus

2.1 Why EDA?

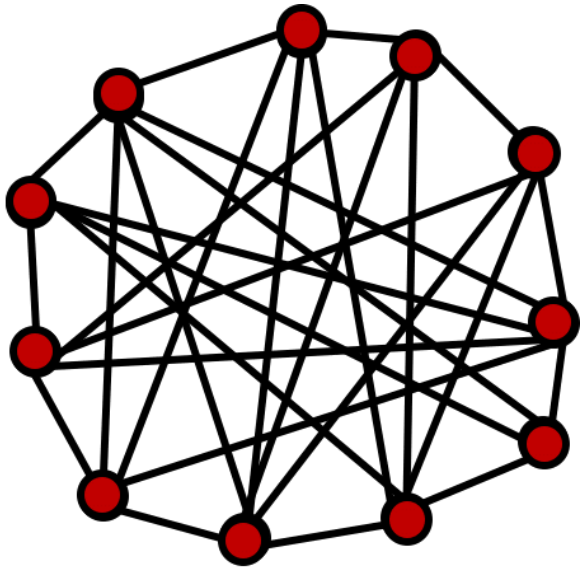
Exploratory Data Analysis ; 탐색적 데이터 분석

가설이 있다! : 출발점이 있음.



"Top-down"

가설이 없다! : 다 뒤져봐야 함.



"Bottom-up"

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

가설을 만드는 첫 번째 원칙 : 관심사를 파악하라. Chapter.1 다시 보기!

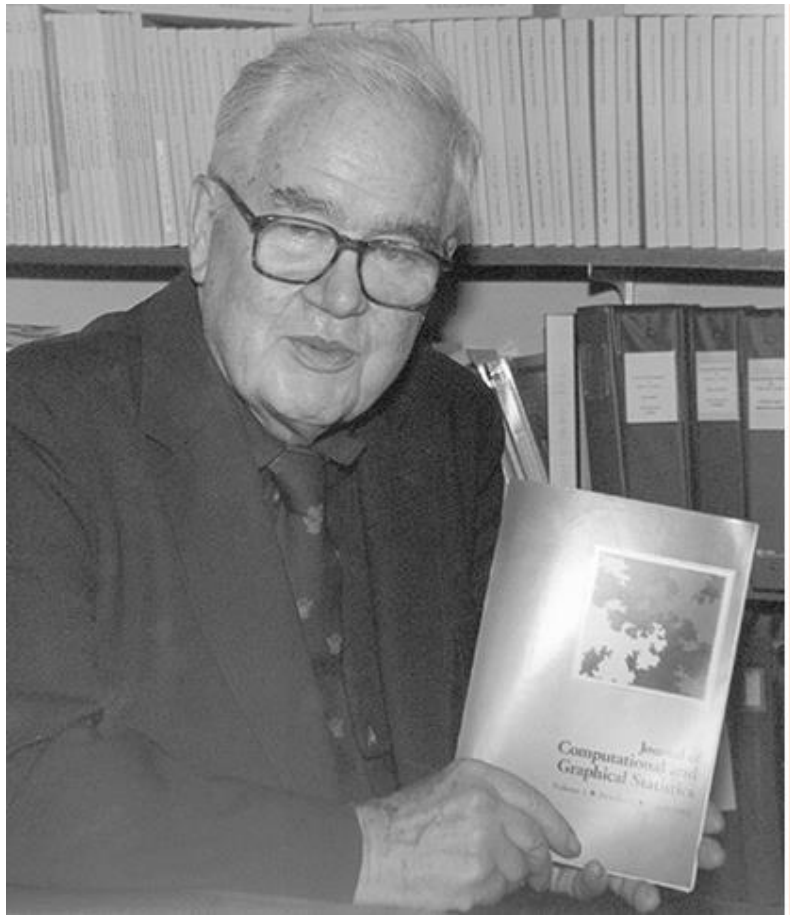
보통 돈 더 벌려면 뭘 해야 할까 고민하면 가설들이 쑥쑥 튀어나옴.

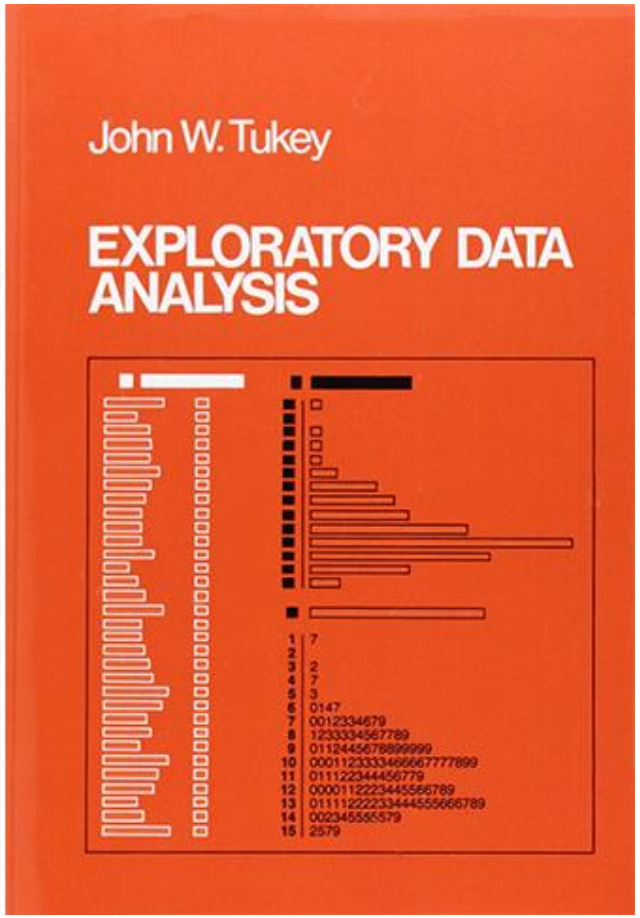
D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

가설을 만드는 두 번째 원칙 : 탐색하라.





2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

탐색적인 데이터 분석을 해야 한다.
현상에 대한 **가설을 세우기** 위해.
가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

탐색적인 데이터 분석을 해야 한다.
현상에 대한 **가설을 세우기** 위해.
가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

데이터 과학, 데이터 마이닝, 빅데이터 분석의 토대가 되는 기술!
어린 학생들에게 통계적 사고 방식을 가르칠 때 사용!

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

한국 이야기는
확실히 아닌 듯 합니다.

탐색적인 데이터 분석을 해야 한다.

가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

데이터 과학, 데이터 마이닝, 빅데이터 분석의 토대가 되는 기술!
어린 학생들에게 통계적 사고 방식을 가르칠 때 사용!

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

어쨌든

탐색적인 데이터 분석을 해야 한다.

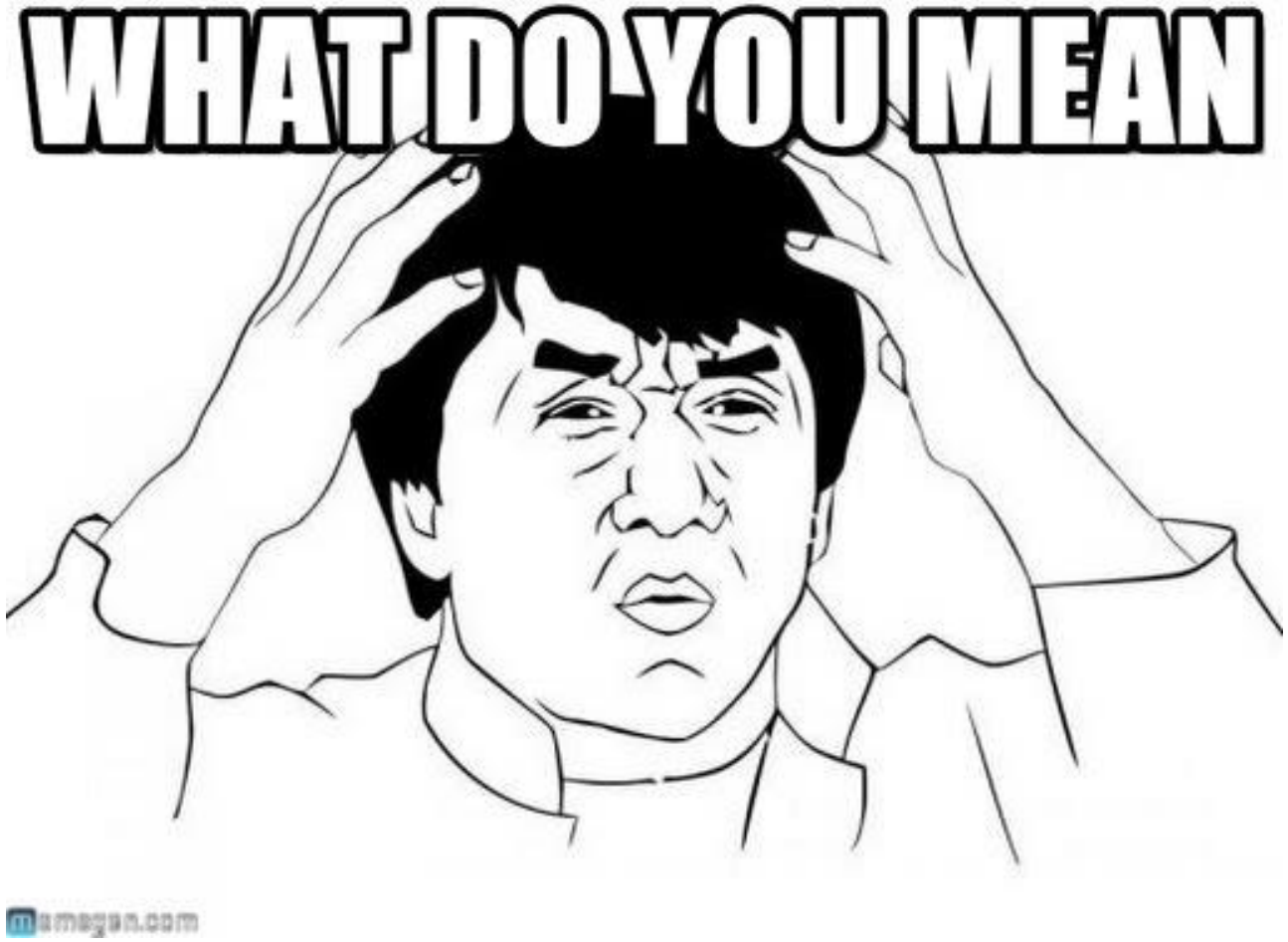
현상에 대한 **가설**을 **세우기** 위해.
가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

데이터 과학, 데이터 마이닝, 빅데이터 분석의 토대가 되는 기술!
어린 학생들에게 통계적 사고 방식을 가르칠 때 사용!

D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석



D:plus

2.1 Why EDA?

Exploratory Data Analysis ; 탐색적 데이터 분석 EASY VER.

가설 검정이고 뭐고
나 뭐 어떻게 분석해야 하는지도 모르겠다니까.

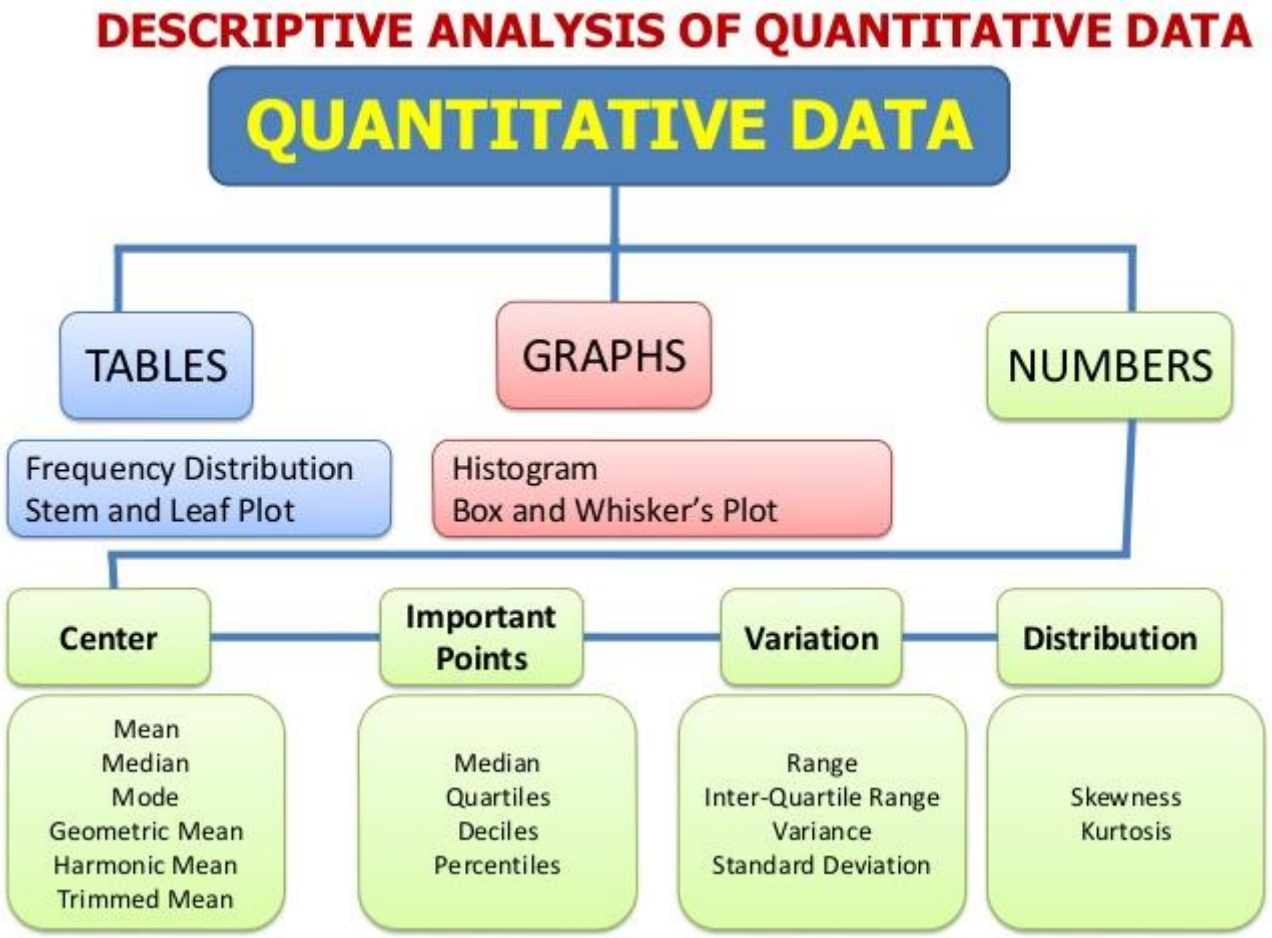
과거에 뭔 현상이 일어났는지 알고 싶음.
머신러닝 머신러닝 말은 많은데 나 이거 써도 됨?
데이터 이거면 충분한 거 아님?
뭐여 우리 데이터 어떻게 생겨 먹은 거냐.

EDA하세요. 길이 보일 겁니다.

D:plus

2.2 4 Types of Questions

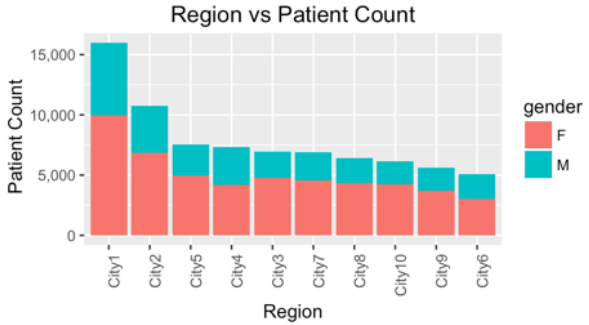
EDA 는 Descriptive Statistics 에 의존합니다.

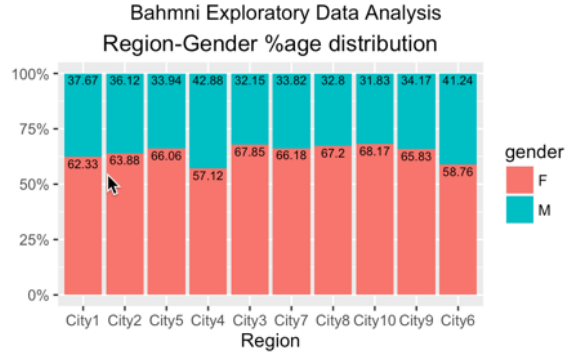


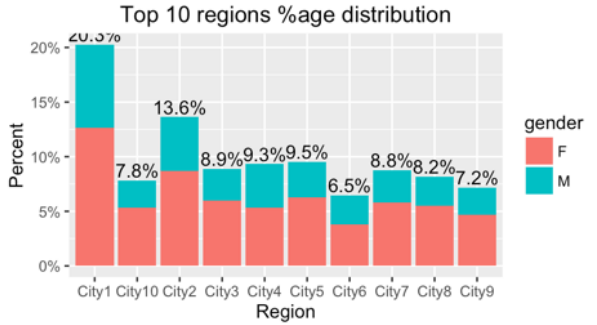
D:plus

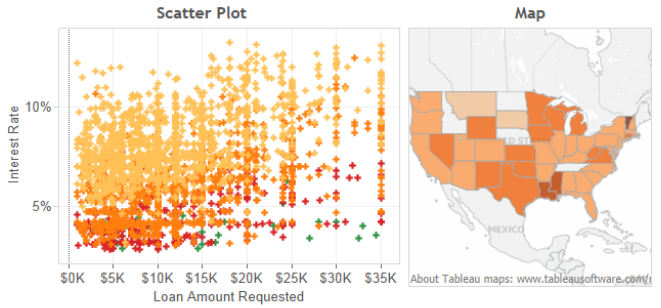
2.2 4 Types of Questions

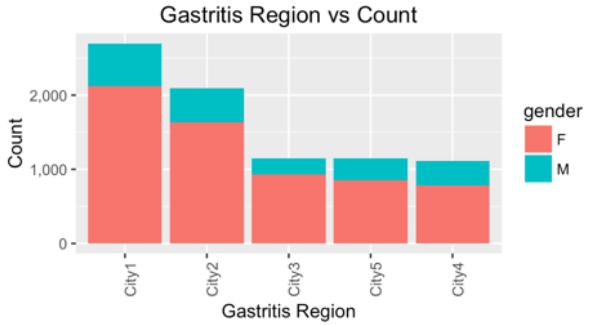
EDA 는 Descriptive Statistics 에 의존합니다.

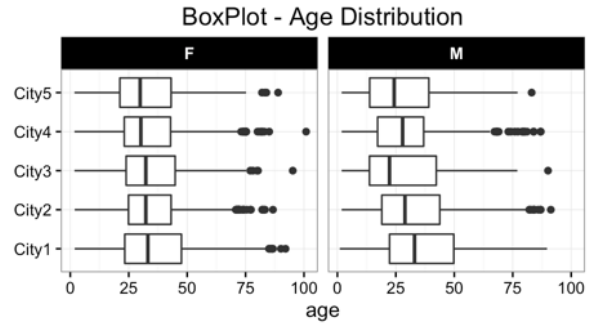


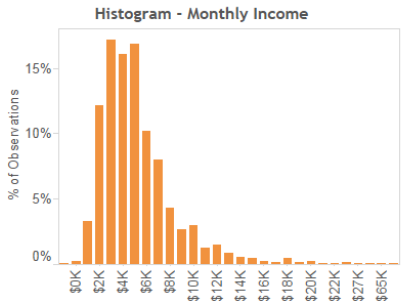


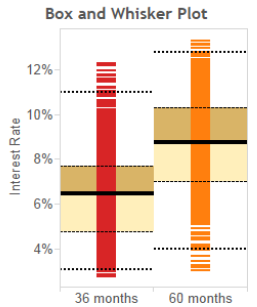


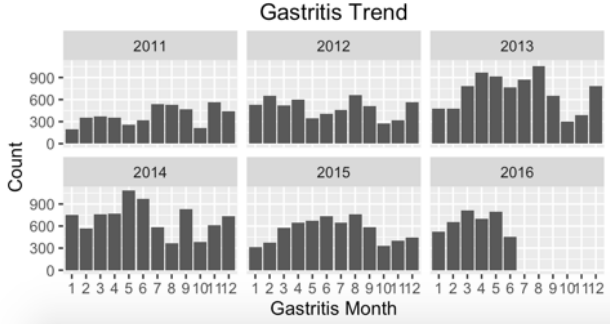


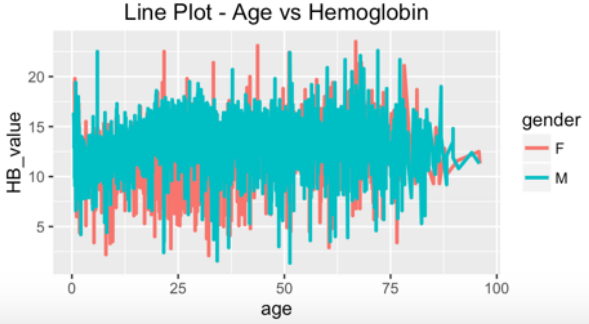


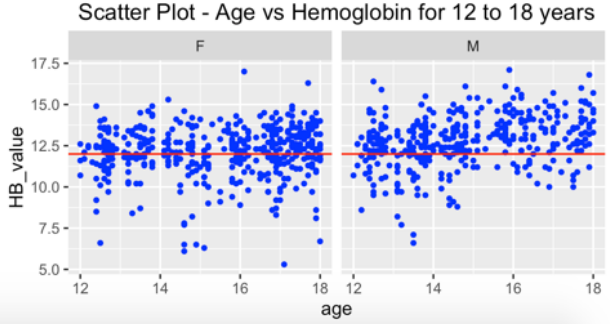








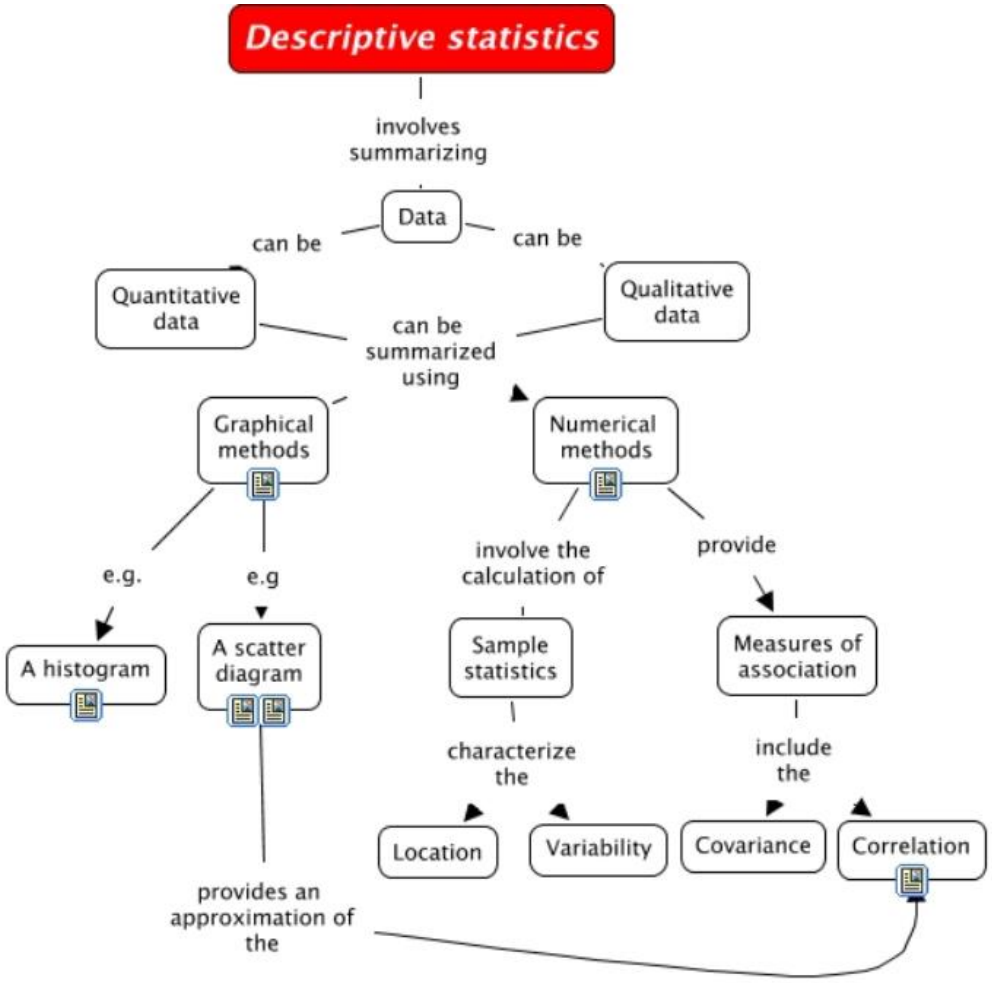




D:plus

2.2 4 Types of Questions

EDA 는 Descriptive Statistics 에 의존합니다.



D:plus

2.2 4 Types of Questions

EDA 는 Descriptive Statistics 에 의존합니다.

무수히 많은 통계의 정의 중 하나 :

통계는,
자료를 요약하고 정리하여
의사 결정에 도움을 주는 수단이다.

위 역할의 첫 출발은 항상 Descriptive Statistics

자료(DATA) = Signal + Noise

Signal을 잡아내는 방법,
[Signal의 크기 측정, Noise의 크기 측정]
등을 위한 온갖 방법론들의 집합.

D:plus

2.2 4 Types of Questions

할 말들 정말 많습니다. 1학기 분량 정도 나와요. **하지만,**

2.2 4 Types of Questions

본질에 집중합시다.

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

1. 현실로 부터.

이 강의에서 첫 번째로 고민하게 되는 것!

자연스럽게 떠오르는 의문 / 가설들을 구체화 하는 방법
의문 / 가설들이 자연스럽게 떠오르게 만들 관점.

D:plus

D:plus

2.2 4 Types of Questions

분석의 선행 조건 1 : 현실의 문제들을 명확히 파악해둘 것.

Q1. 무엇이 목표일까요?

Q2. 무슨 문제들이 있을까요?

Q3. 불편한 점 없으신가요?

D:plus

2.2 4 Types of Questions

자, 여러분은 레모네이드 트럭 사장님이십니다

데이터 수집과 실험을 중요하게 여기시는 모습이시다.

D:plus

2.2 4 Types of Questions

분석의 선행 조건 2 : 현실에 맞추어 데이터를 확인할 것

Date	Location	Lemon	Orange	Temperatu	Leaflets	Price
7/1/2016	Park	97	67	70	90	0.25
7/2/2016	Park	98	67	72	90	0.25
7/3/2016	Park	110	77	71	104	0.25
7/4/2016	Beach	134	99	76	98	0.25
7/5/2016	Beach	159	118	78	135	0.25
7/6/2016	Beach	103	69	82	90	0.25
7/6/2016	Beach	103	69	82	90	0.25
7/7/2016	Beach	143	101	81	135	0.25
	Beach	123	86	82	113	0.25
7/9/2016	Beach	134	95	80	126	0.25
7/10/2016	Beach	140	98	82	131	0.25
7/11/2016	Beach	162	120	83	135	0.25
7/12/2016	Beach	130	95	84	99	0.25
7/13/2016	Beach	109	75	77	99	0.25

D:plus

2.2 4 Types of Questions

질문 / 아이디어 / 가설을 구체화 할 4가지 관점을 소개합니다.

- **Descriptive type**
- **Associative type**
- **Comparative type**
- **Predictive type**

가설을 구체화 시키는 방법론으로도, EDA 순서로도 사용 가능!

D:plus

2.2 4 Types of Questions

Descriptive type

얼마나 팔았지?

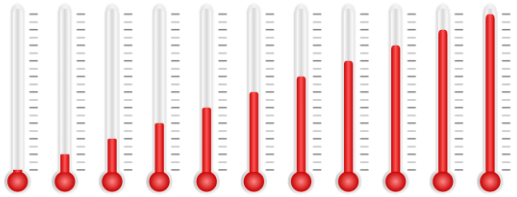
관심사 : 레몬에이드 판매량

1. 판매량에 대한 궁금증들을 마구마구 꺼낸다.
2. 판매량 데이터만 가지고 관찰할 수 있는 질문들만을 추린다.
3. 질문에 대한 적절한 통계 방법을 선택한다.

D:plus

2.2 4 Types of Questions

Associatvie type



관심사 : 레몬에이드 판매량

1. 판매량에 대한 궁금증들을 마구마구 꺼낸다.

2. 다른 무언가로 판매량을 설명하려는 궁금증들만을 남긴다.

3. 질문에 대한 적절한 통계 방법을 선택한다.

2.2 4 Types of Questions

Comparative type

레몬에이드, 오렌지에이드.

어떤 것이 더 잘 팔리지?

관심사 : 이익

1. 판매량에 대한 궁금증들을 마구마구 꺼낸다.
2. 선택의 문제가 되는 경우를 추린다.
3. 적절한 통계 방법을 선택한다.

2.2 4 Types of Questions

Predictive type

여러 욕망과 분노, 기대가 뒤섞인 질문

얼마나 팔릴까?!

2.2 4 Types of Questions

사실, 칼 같이 나뉘지는 않습니다.

Descriptive Type.

>현상에 대한 관찰이 필요할 때. 요약이 필요할 때.
>주로 과거 데이터를 살펴볼 때
>관심사를 설명하기 위한
  기초적인 구조를 잡을 때

Associative type
Comparative type

> 미래에 어떻게 변할지 알고 싶을 때

Predictive Type.

2.2 4 Types of Questions

Summary

- 현실에서 무슨 문제를 풀어야 하는지 명확히 한다. 최소한, 현실에 공감하고 상상할 수 있는 상황이어야 한다.
- 데이터를 보고 관심사에 맞추어 질문들을 이끌어낸다. 마인드맵처럼 쭉쭉 뽑아내도 좋다.
- 데이터에 없는 질문도 좋다. 그 데이터가 있다면 뭐가 좋은지 구체적으로 상상하라. [새로 수집해야 하는 데이터]가 된다.
- 처음에는 관심사를 먼저 요약해본다. 많은 불편함을 줄 것이다. 이는 가설과 아이디어를 위한 밑바탕이 된다. Question : [ ] type
- [관심사를 다른 무언가로 설명할 수 있을까?] Question : [ ] type
- 선택의 문제가 될 경우 [ ] type이다. 비교는 정말 강력한 도구다.
- Predictive type 1 : A를 조작하면 B가 이만큼 올라갈까?
- Predictive type 2 : 앞으로 A가 이렇게 바뀌게 될 거야. 대비해야 해.

D:plus

2.2 4 Types of Questions

주의 사항. 순환 논증에 빠지지 말자.

Testing Hypotheses Suggested by the Data

데이터를 관찰하고 가설을 끌어냄
끌어낸 가설을 방금 전 그 데이터를 바탕으로 검정 함.
그 가설이 맞게 나옴 (대부분 그렇게 될 수 밖에 없음)

올바른 방식은?

EDA를 통해 나온 가설을, 그 데이터 위에서 검정했다
→ 그 검정결과는 **[가설에 대한 채택/기각]**의 문제가 아니라, **[가설 자체가 그럴 법 한지]**에 대한 원천이 된다.
→ 검정 결과 **[가설이 맞든, 틀리든]** 반드시 가설을 더 구체화하여 **새로운(미래의)데이터 위에서 확인해봐야 한다**. (실험계획)

2.2 4 Types of Questions

백문이 불여일견.

실습하러 갑시다.