Mid. Project in R

통합 실습 시나리오

실습목표	▪ 비즈니스 이해 단계에서 데이터 준비단계 까지의 절차를 이해하고 데이터 분석을 수행할 수 있다.
통합실습 시나리오 (미션)	**주가 예측** 1. SK 주식회사의 4년치 주가 데이터가 주어진다. 2. 2016년 1월 ~ 2019년 11월까지 데이터를 학습해서 2019년12월 데이터로 검증(평가)하고자 한다. 3. 전날 장이 종료된 이후 다음날 종가를 예측하는 모델을 만든다. 4. 이에 맞게 가설을 수립하고, 데이터를 구성(전처리)하고, 탐색적 분석과 가설검증을 수행한다. 5. 기본 모델링을 통해 예측 모델을 만들고 분류 문제의 평가를 수행한다.
진행방식	▪ 총 5개의 과제로 구성 - 과제1. 데이터 둘러보기 - 과제2. 가설수립 - 과제3. 데이터 전처리 - 과제4. EDA - 과제5. 기본 모델링 및 평가 ▪ 각 과제마다 다음의 절차로 진행 : 팀 수행 > 제출 > 개인 발표 > 풀이

주의사항

- ✓본 과정의 목표는 지금까지 배워온 내용을 총 복습하는 것입니다.
- ✓그래서 이론 설명이 부족하다고 느끼실 수 있습니다.
- ✓본 과정은 평가를 진행하지는 않습니다.
- ✓본 과정에서 주가 예측 정확도는 가설수립과 EDA를 통한 중요 변수 발굴에 달려 있습니다.
- ✓혹시라도 주가 예측 결과에 고무되어 직접 투자하는데 사용하시면 절대 안됩니다!!!!!

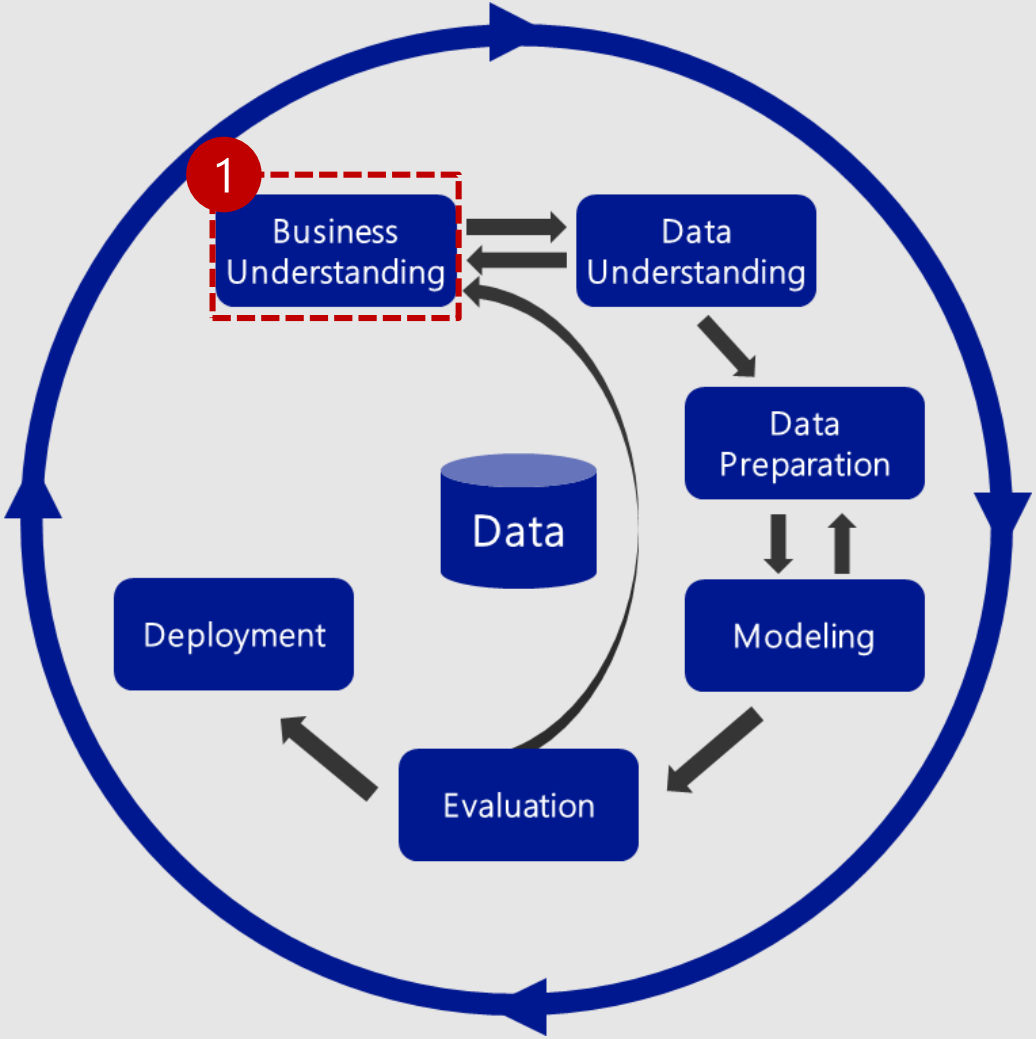
- ✓특전 : 주가 예측 성능이 가장 좋게 나온 팀에게는 ##벅스 커피쿠폰을 드립니다!

데이터분석 표준 프로세스 Review

CRISP-DM

- ✓ 프로세스를 한번 거쳤음에도 문제가 해결되지 않을 수 있다
  ➔ 그렇다고 실패가 아님!

- ✓ 한번에 해결책을 찾지 못해도 데이터를 더 잘 이해하게 되는 계기가 됨
  ➔ 두번째 수행할 때는 더 많은 정보를 갖고 시작할 수 있음!



①Business Understanding

✓개요

- 잘 정의된 명확한 데이터분석 문제로 시작하는 프로젝트는 거의 없음.
- 문제를 파악해 가는 과정을 반복하면서 문제를 재정의하고 해결책을 정의하게 됨.

✓수행되는 내용

- 비즈니스 목표 검토
- 데이터 분석 목표 수립
- (초기)가설 수립

①Business Understanding

✓비즈니스 목표에서 데이터 분석 목표로…

비즈니스 관점	**목표**	유통 매장 매출 100억 달성 (작년실적 70억)
	문제점	고객을 경쟁 매장에 빼앗기고 있음.
	방법	이탈할 것으로 보이는 고객을 붙잡기 위한 마케팅
데이터 분석 관점	**문제정의**	고객이 이탈할 지 사전에 예측할 수 있을까?
	목표	✓ 이탈고객 정의 : 3개월간 매장에 방문하지 않는 고객 ✓ 어느정도 정확도로 예측을 하면, 매출목표 달성에 기여할 수 있을까?
	분석	▪ 분류문제 ▪ **고객 이탈**에 영향을 미치는 **요인**은 무엇일까?

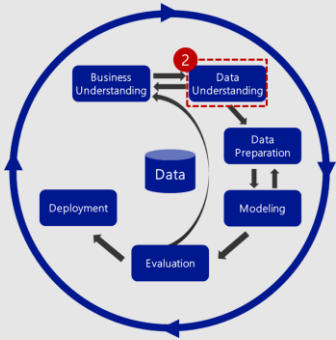
①Business Understanding

✓(초기)가설 수립

고객 이탈에 영향을 미치는 **요인**은 무엇일까?

- 다양한 직무에 있는 사람들의 의견을 수렴할 필요가 있음.
- 데이터의 존재여부를 고려하지 말고 가설 도출.
- 초기 가설 수립 이후 데이터 탐색을 통해 가설을 구체화

②Data Understanding



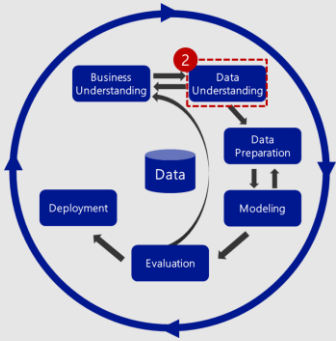
✓개요

- 데이터 : 문제의 해결책을 만드는 데 사용할 원자재
- **문제에 정확히 부합하는 데이터가 있는 경우는 거의 없음.**
- 데이터에 따라 데이터 취득 및 유지 비용이 다름.

✓수행되는 내용

- 데이터 원본 식별 및 취득
- 데이터 탐색 : EDA, CDA

②Data Understanding



✓ 데이터 원본 식별 및 취득

- (초기)가설에서 도출된 데이터의 원본을 확인

있는 데이터

없는 데이터

가용한 데이터

그대로 사용가능

가공해야 사용가능

취득 가능

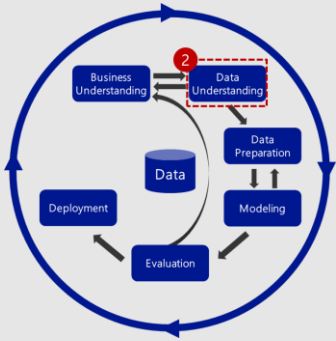
취득 불가능

Data Pipeline
Data Environment
검토 및 과제 수행

취득 비용 산정
과제 수행

정보의 의미 정의
→ 정보 분할
→ 최대한 가용한 데이터 영역으로 만들기

②Data Understanding



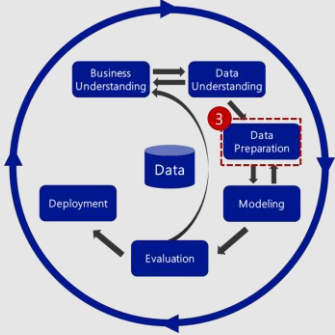
✓데이터 탐색 : EDA, CDA

- 데이터를 탐색하는 두 가지 방법

데이터 통계량	데이터 시각화
분할표(Contingency Table) MIN, MAX, SUM, MEAN Quartile …	Histogram, Box plot, Density plot Bar plot, Pie chart Scatter plot …

- EDA (Exploratory Data Analysis)
  - 개별 데이터의 분포, 가설이 맞는지 파악
  - NA, 이상치 파악
- CDA (Confirmatory Data Analysis)
  - 탐색으로 파악하기 애매한 정보는 통계적 분석 도구(가설 검정) 사용

③ Data Preparation



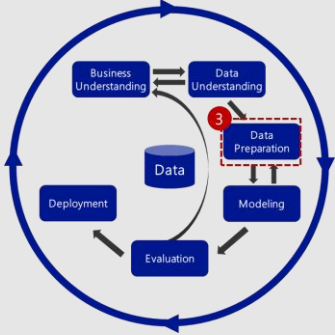
✓개요
- 데이터 분석을 위해 특정 조건에 맞는 데이터 유형과 구조가 있음
- 더 좋은 결과를 얻을 수 있도록 데이터의 형태를 조작하고 변환하는 과정 필요.

✓수행되는 내용
- 데이터 정제
- 추가 변수(Feature Engineering)

✓결과물 : **하나의 잘 정리/정제된 데이터프레임(테이블)**

(주의) 데이터 준비 단계가 끝나면, 사실 EDA를 다시 수행하며 데이터를 확인 합니다.

③ Data Preparation

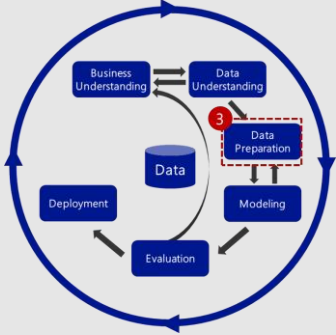


✓ 데이터 정제

- 잘못된 데이터 정제
- 결측치(NA) 식별 및 조치
  - 중요한 요인에 결측치가 존재한다면 반드시 조치해야 한다.
  - 예 : 옷을 추천하는데, 고객의 나이나 성별에 결측치가 존재한다면, 옷을 추천하기 곤란.

- 이상치 식별 및 조치
  - 잘못된 값
  - 값 자체는 정상이나 다른 값들의 분포에 비해 치우친 값
  - 이러한 값은 데이터 분석 시 잘못된 결과를 얻게 하는 원인이 됩니다.

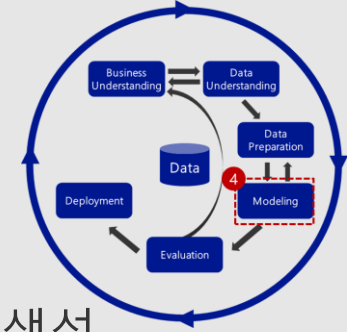
③ Data Preparation



✓ 추가변수(Feature Engineering)

- 기존에 저장된 데이터를 그대로 사용해서는 제대로 된 예측 결과를 얻기 어렵다.
- 데이터베이스에 데이터를 저장하는 방식
  - 트랜잭션 발생 순으로 저장 ➔ 저장된 데이터 자체가 비즈니스의 Insight가 되지 못함.
- 비즈니스의 경험 + 데이터 분석을 통해 인사이트를 발견하고, 이를 담아내는 정보가 필요
- 사례
  - 페이스북 고객 중 **가입 후 10일 이내 7명의 친구를 사귄 사람**은 그렇지 않은 사람보다 잔존율이 훨씬 높다!
  - 음주 습관에 대한 분석 : age 변수를 이용해서 age >= 20 ➔ 음주가능연령
  - 아파트가격 분석 : 방 수 >=4 & 화장실수 >=2 ➔ Premium

④ Modeling



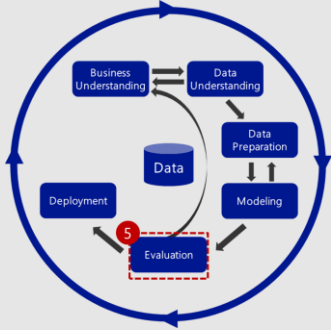
✓개요

- 중요 변수들을 선택하고, 적절한 알고리즘을 적용하여 예측 모델을 생성
- 생성된 모델을 평가

✓수행되는 내용

- 데이터셋 분리
- 중요 변수 선정
- 머신러닝 알고리즘 적용하여 모델 생성
- 모델 테스트

⑤ Evaluation



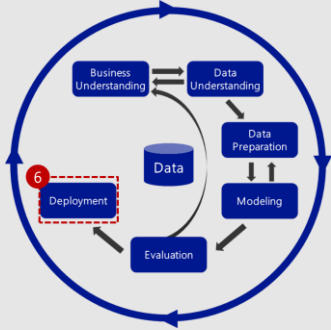
✓개요

- 모델에 대한 데이터 분석 목표와 비즈니스 목표달성에 대한 평가
- 모델과 데이터에서 추출한 패턴이 진정한 규칙성을 갖고 있는지, 단지 특정 예제 데이터에서만 볼 수 있는 특이한 성질은 아닌지 확인
- 비즈니스 목표에 부합되는지 보장

✓수행되는 내용

- 모델에 대한 최종평가 : Test Set 이용
- 비즈니스 기대가치 평가

⑥ Deployment



✓개요

- 프로젝트 결과물 최종 확정: 프로덕션 환경의 파이프라인, 모델 및 배포가 고객 목표를 충족하는지 확인
- 운영시스템에서 품질(성능 목표) 유지 기준을 정하고, 모니터링 계획을 수립

✓수행되는 내용

- 시스템 유효성 검사: 배포된 모델과 이 고객 요구 사항을 충족 하는지 확인
- 프로젝트 이전 : 운영환경으로 배포

Mission #1 : 데이터 둘러보기

✓'과제1.데이터 둘러보기.R' 안의 문제를 수행하시오.
✓주가 데이터, 환율 데이터가 주어집니다..

SK
주가

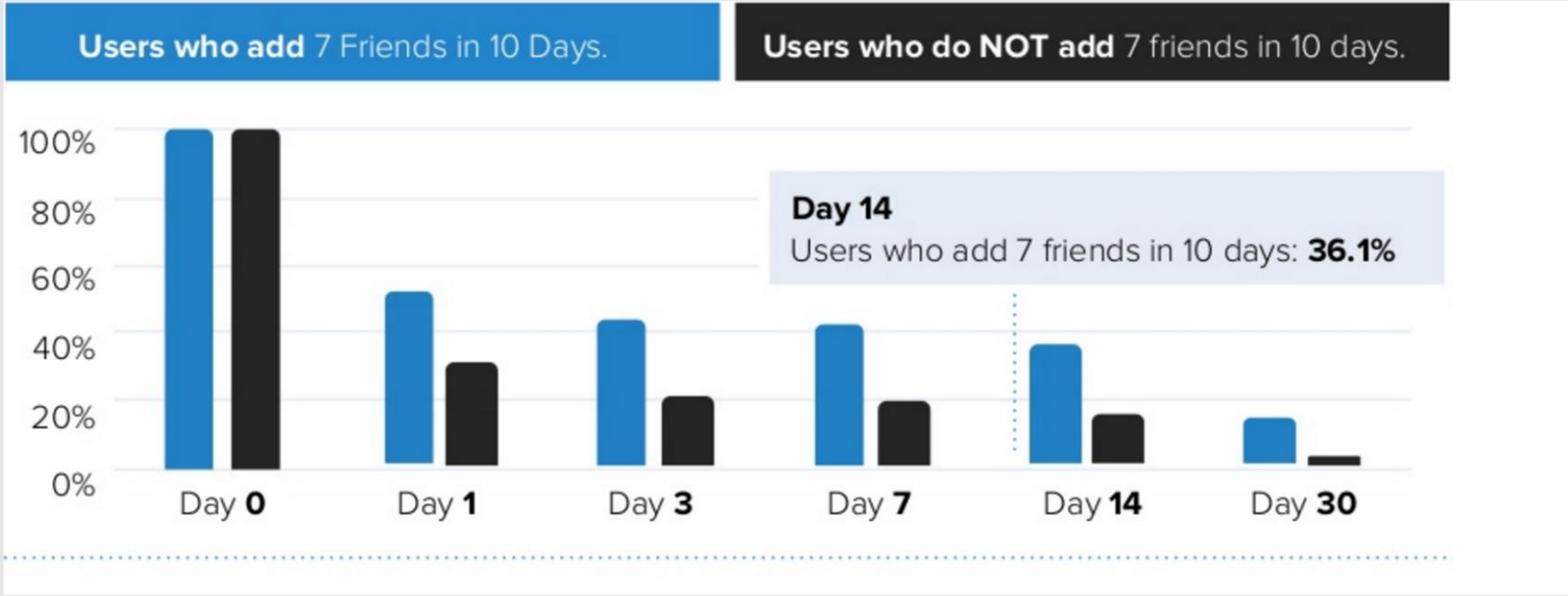
	Date	Open	High	Low	Close	Adj Close	Volume
2	2016-01-04	243000.000000	245000.000000	234500.000000	234500.000000	226585.656250	173905
3	2016-01-05	236000.000000	244000.000000	234000.000000	241000.000000	232866.281250	182985
4	2016-01-06	241000.000000	243000.000000	237500.000000	239000.000000	230933.781250	108574
5	2016-01-07	237000.000000	243000.000000	236000.000000	240500.000000	232383.156250	113376
6	2016-01-08	240500.000000	242500.000000	235000.000000	241500.000000	233349.406250	81557
7	2016-01-11	238000.000000	241500.000000	236000.000000	239000.000000	230933.781250	84152
8	2016-01-12	240000.000000	246000.000000	237000.000000	237500.000000	229484.406250	86196

원-달라
환율

date	close	open	high	low	diff
2019-12-31	1,155.07	1,157.97	1,159.19	1,153.14	-0.0025
2019-12-30	1,157.97	1,160.97	1,161.30	1,154.89	-0.0015
2019-12-27	1,159.68	1,162.15	1,162.83	1,158.12	-0.0023
2019-12-26	1,162.30	1,161.04	1,163.42	1,160.34	0.0013
2019-12-25	1,160.76	1,163.52	1,164.09	1,160.68	-0.0024
2019-12-24	1,163.52	1,164.04	1,166.61	1,160.71	-0.0004

가설 수립

(핵심)가설을 찾아서...


Users who add 7 Friends in 10 Days.
Users who do NOT add 7 friends in 10 days.
100%
80%
60%
40%
20%
0%
Day 14
Users who add 7 friends in 10 days: 36.1%
Day 0
Day 1
Day 3
Day 7
Day 14
Day 30

(핵심)가설을 찾아서...

여러분이, SNS를 운영 중이라고 가정해 봅시다.

데이터 분석팀

10일 이내에 친구 7명을 사귄 사용자는 우리 서비스를 오래 이용해주더군요!

사실 이런 종류의 문장은
수도 없이 찾아낼 수 있습니다.

(핵심)가설을 찾아서…

✓ **고객 행동 관점의** Business Model

✓ 매 단계가 관심사이나, 실제 매출은 'Facebook**을 계속 사용하는 고객**'으로부터 발생

✓ 그렇다면 관심사를 좀 더 **구체적**으로 말하면 ➔ '**고객의 잔존(율)**'

(핵심)가설을 찾아서

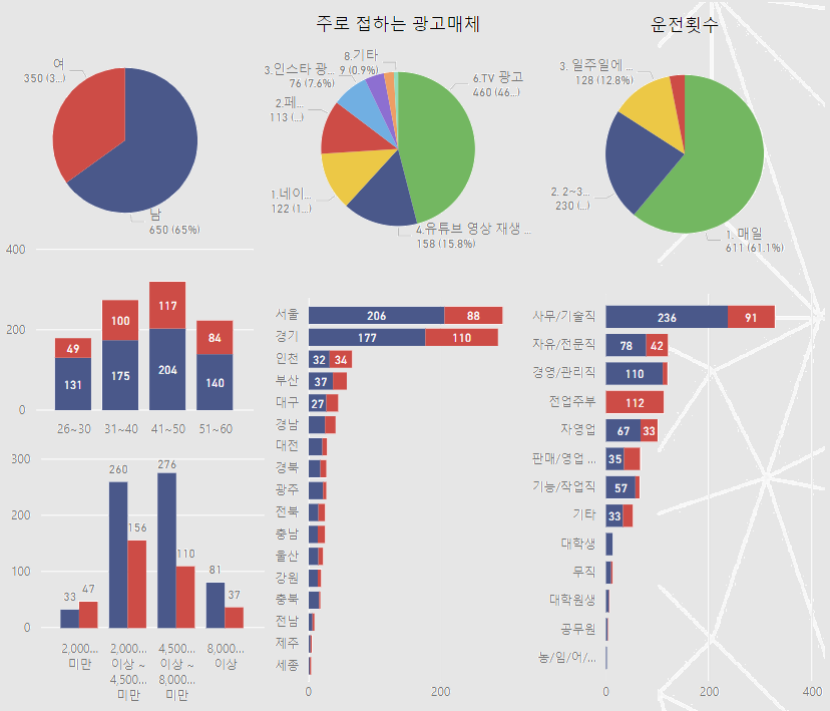
✓Summary

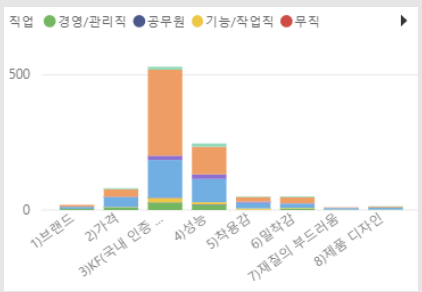
① 분명한 [관심사], [목표]를 찾아라.

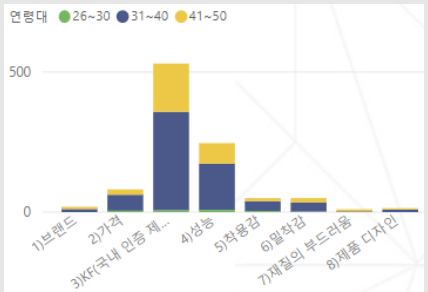
(핵심)가설을 찾아서...

✓찾아낸 [목표]에 영향을 주는 정보가 무엇인지 탐색합니다.

- 샘플 데이터를 이용 : 제한된, 당장 사용 가능한 데이터 이용
- 데이터를 탐색 : R, Python, SQL, Excel, Power BI 등 사용가능한 도구 활용.







Q08_1순위	Q12_0.지인에게 추천한 적 없음	Q12_1.가격이 저렴해서	Q12_2.기능이 좋게 느껴져서	Q12_3.착용감이 좋아서	Q12_4.디자인이 좋아서	Q12_5.가성비가 좋아서	Q12_6.신뢰하는 브랜드여서	Q12_7.기타(직접 입력)
1)브랜드	11	4	2	2	1	3	3	0
2)가격	32	19	20	10	3	29	9	0
3)KF(국내 인증 제품) 등급	211	76	157	102	19	174	78	3
4)성능	99	39	89	43	16	77	26	0
5)착용감	12	13	18	17	8	19	9	0
6)밀착감	13	9	18	23	6	10	9	0
7)재질의 부드러움	2	3	4	2	1	2	1	0
8)제품 디자인	4	0	3	1	4	8	2	0
합계	384	163	311	200	58	322	137	3

(핵심)가설을 찾아서...

✓목표에 대한 질문 : Facebook고객을 오래 잔존하게 하려면?

✓지금 가지고 있는 정보

- 성별
- 나이
- 가입일
- 방문 횟수
- 포스팅 수
- 좋아요 누른 수
- 좋아요 눌러준 수
- 덧글 단 수
- 덧글 달아준 수
- 친구 수
- 커뮤니티 가입 수

✓위 정보를 가지고 어떻게 탐색을 하면 좋을까요?

(핵심)가설을 찾아서...

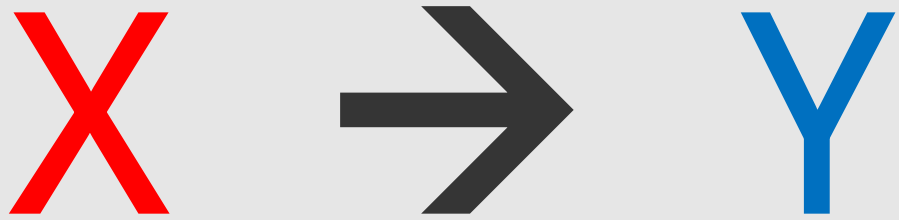
✓Summary

① 분명한 [관심사], [목표]를 찾아라.

② [관심사], [목표]를 설명할 요인을 찾아라.

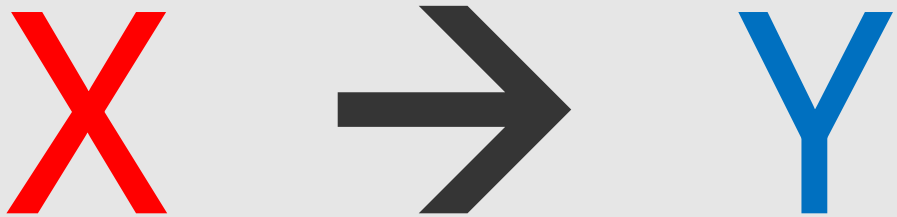
가설 구조

10일 이내에 친구 7명을 사귄 사용자는
우리 서비스를 오래 이용해주더군요!



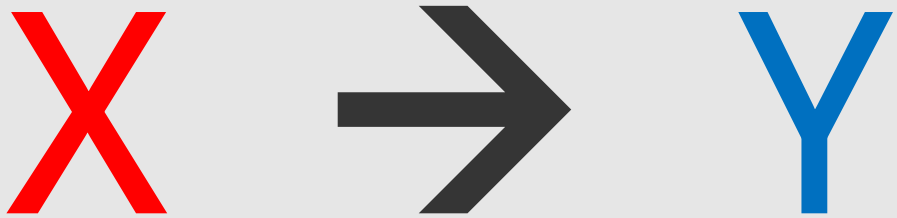
X : 10일 이내 친구 7명 이상 여부
Y : 잔존율
→: 차이가 날 것이다.

가설 구조



X : 어제 뿌린 광고지의 개수
Y : 아이스크림 판매량
→: 비례할 것이다.

가설 구조



X : **미세먼지 농도**
Y : **황사마스크 판매량**
→: 비례할 것이다.

그런데, 우리가 X를
제어할 수 있는가?

(핵심)가설을 찾아서...

✓Summary

① 분명한 [관심사], [목표]를 찾아라.

② [관심사], [목표]를 설명할 요인을 찾아라.

③ 가설의 구조를 정의하라. X → Y
단, 결과(Y)에 대해 요인(X)를 제어할 수 있어야 가설이다.

Mission #2 : 가설 수립하기

✓ '과제2.주가_예측_가설수립' 안의 문제를 수행하시오.

- '가설도출' 시트를 채우시오.
- '데이터셋 구조' 에 가설을 붙여서 데이터셋 구조가 이상하면 가설을 수정해야 합니다.
- 주어진 데이터 외의 데이터(정보)를 인터넷에서 다운받아 활용하셔도 됩니다.
  - 예 : KOSPI 지수, 미국 다우존스 지수 , 경기 동향 지표 등

데이터 준비

Mission #3 : 데이터 준비

가설로 도출된 요인들과 주가를 하나의 데이터프레임으로 만드시오.

- 단, 한 행은 일별 주가 이어야 합니다.

Date	주가	x1	x2	...	xn

탐색적 데이터 분석

탐색적 데이터 분석

✓ Exploratory Data Analysis
- 통계와 그래프를 이용해서 대상 데이터를 파악하는 것.
- 본격적인 분석에 들어가기 전에 반드시 거쳐야 할 단계

✓ EDA**를 통해 무엇을 파악해야 하는가**?

① 각 변수들의 분포(결측치, 이상치 포함)
② 종속변수(Target)와 Feature들의 관계
③ 변수들 간의 관계

Mission #4 : 탐색적 데이터 분석

- ✓ 개별 변수의 분포를 살펴봅니다.
- ✓ 종속변수(Target)과 Feature들과의 관계를 살펴봅니다.
- ✓ Target과 관련이 높은 변수들, 낮은 변수들, 애매한 변수들로 구분해 봅시다.
- ✓ 추가로 Feature들 끼리의 관계도 살펴봅시다.

Mission #5 : 가설 검증

- ✓ Mission #4 EDA 의 2단계에 대해서 모두 가설 검증을 수행하시오.

모델링

Mission #6 : 모델링 및 평가

본 과정은 모델링에 초점을 두지 않으므로 기본 코드가 제공됩니다.

- ✓ 데이터 셋은 2019년 11월까지의 데이터와 2019년 12월 데이터를 분리합니다.
- ✓ 선형 회귀분석 알고리즘을 이용하여 모델을 생성합니다.
- ✓ 예측 결과를 평가(mae, mape)하고
- ✓ 시각화해서 비교해 봅시다.