December 10, 2025

▣ **강의명:** CS109A: 데이터 과학 입문

▣ **주차:** Lecture 20

▣ **교수명:** Pavlos Protopapas, Kevin Rader, Chris Gumb

▣ **목적:** Lecture 20의 핵심 개념 학습

Contents

□ 핵심 요약

이 문서는 데이터 분석의 두 가지 큰 장애물인 **결측치(Missingness)**와 **블랙박스 모델(Black Box Models)**을 다루는 방법을 설명합니다.
1. **결측치 처리:** 데이터에 구멍(NaN)이 있을 때 발생하는 문제를 이해하고, 단순 삭제나 평균값 대체의 위험성(편향)을 배웁니다. MCAR, MAR, MNAR이라는 세 가지 결측 유형을 구분하고, '결측치 표시자'나 '모델 기반 대체'(특히, 불확실성을 고려한 대체)와 같은 고급 기법들을 살펴봅니다. 2. **모델 시각화:** k-NN이나 의사결정나무처럼 해석이 어려운 '블랙박스' 모델의 작동 방식을 이해하기 위해 **부분 의존성 플롯(Partial Dependence Plots, PDP)**을 사용하는 방법을 배웁니다. PDP를 통해 특정 변수가 모델의 예측에 어떤 영향을 미치는지, 변수 간 상호작용은 없는지 시각적으로 확인할 수 있습니다.

1 용어 정리

주요 용어들을 표로 정리했습니다.

용어	원어	쉬운 설명	비고
결측치	Missingness	데이터셋에 값이 누락된 '구멍'이 있는 상태 또는 그 값.	`NaN` (Not a Number)로 표시됨.
편향	Bias	모델의 예측이나 추정치가 체계적으로 한쪽으로 치우치는 경향.	예: 저울이 항상 500g 높게 측정하는 것.
대체 (대치)	Imputation	누락된 결측치를 추정된 값으로 채워 넣는 과정.	가장 간단한 예는 '평균값'으로 채우기.
MCAR	Missing Completely at Random	**완전 임의 결측.** 결측이 다른 어떤 변수와도 상관없이 무작위로 발생.	(예: 데이터 입력 중 무작위 오타)
MAR	Missing at Random	**임의 결측.** 결측 여부가 *다른 관측된* 변수와 관련 있음.	(예: 남성이 여성보다 '소득' 문항 응답률이 낮음)
MNAR	Missing Not at Random	**비임의 결측.** 결측 여부가 *누락된 값 자체* 또는 *관측되지 않은* 변수와 관련 있음.	(예: 고소득자가 '소득' 문항 응답을 거부함)
블랙박스 모델	Black Box Model	모델의 내부 작동 원리를 이해하기 어려운 복잡한 모델.	예: k-NN, 랜덤 포레스트, 딥러닝
PDP	Partial Dependence Plot	**부분 의존성 플롯.** 다른 변수들을 고정한 채, 특정 변수 하나가 모델 예측에 미치는 영향을 보여주는 그래프.	모델 해석의 핵심 도구.

Table 1: 결측치 및 모델 해석 관련 핵심 용어

2 결측치(Missingness)란 무엇인가?

2.1 결측치의 정의와 문제점

결측치(Missingness)란 데이터셋에 값이 존재하지 않는 '구멍'이 있는 것을 의미합니다. Pandas에서는 주로 `NaN` (Not a Number)로 표시됩니다.

데이터에 결측치가 있으면 두 가지 큰 문제가 발생합니다.

1. **모델 학습 불가:** 대부분의 머신러닝 라이브러리(예: `sklearn`)는 데이터에 `NaN` 값이 하나라도 포함되어 있으면 오류를 발생시키며 모델 학습을 거부합니다. 2. **심각한 편향(Bias) 발생:** 결측치를 처리하기 위해 순진한 방법을 사용하면, 모델이 현실을 체계적으로 잘못 예측하는 '편향'이 발생할 수 있습니다.

주의사항

편향(Bias)이란?
편향은 모델의 예측이 지속적으로 한쪽으로 치우치는 것입니다.
* **직관적 예시:** 어떤 저울이 항상 실제 무게보다 1kg을 더 높게 보여준다면, 이 저울은 '편향'되어 있습니다. 몇 번을 측정하든 항상 1kg 높은 값이 나옵니다. * **데이터 예시:** 결측치를 단순히 '0'으로 채웠다고 가정해봅시다. 만약 '0'이라는 값이 실제 데이터에서는 매우 드문 값이라면, 모델은 '0'이라는 값을 과도하게 학습하여 현실과 동떨어진 예측을 하게 될 수 있습니다.

2.2 결측치 처리의 순진한 접근법 (과 그 위험성)

결측치를 만났을 때 가장 먼저 떠올리는 간단한 두 가지 방법과 그 위험성은 다음과 같습니다.

1. **관측치(행) 삭제:** 결측치가 하나라도 있는 행(row)을 모두 삭제합니다. * **위험성:** 만약 결측치가 무작위로 발생한 것이 아니라 특정 그룹(예: 특정 연령대)에서만 집중적으로 발생했다면, 해당 그룹의 데이터가 통째로 사라집EBF니다. 이는 모델이 해당 그룹을 아예 학습하지 못하게 만들어 심각한 편향을 유발합니다. 2. **단순 값 대체 (Mean/Median/Mode):** * **수치형 변수:** 전체 데이터의 '평균(mean)'이나 '중앙값(median)'으로 모든 `NaN`을 채웁니다. * **범주형 변수:** 가장 빈번하게 등장한 '최빈값(mode)'으로 모든 `NaN`을 채웁니다. * **위험성:** 모든 결측치에 똑같은 값을 넣으면, 해당 값 주변에 데이터가 비정상적으로 몰리게 됩니다. 이는 데이터의 실제 분포를 왜곡하며, 변수 간의 관계를 약화시키거나 왜곡시킵니다.

이러한 순진한 방법들은 데이터의 귀중한 정보를 손실시키고 편향을 유발하므로, 왜 결측치가 발생했는지 먼저 파악하는 것이 중요합니다.

3 결측의 3가지 유형 (MCAR, MAR, MNAR)

결측치를 제대로 처리하기 위해서는 결측치가 "왜" 발생했는지 그 원인을 파악해야 합니다. 통계학에서는 이 원인을 세 가지 유형으로 분류합니다.

3.1 1. MCAR (Missing Completely at Random, 완전 임의 결측)

"결측이 완전히 무작위로 발생했다"는 의미입니다.

이는 결측 여부가 데이터셋의 그 어떤 변수(관측된 변수, 누락된 변수)와도 아무런 관련이 없는 경우입니다.

* **비유:** 설문지 데이터를 엑셀에 입력하다가, 직원이 피곤해서 아무 데나 랜덤하게 몇 개를 빠뜨리고 입력한 상황입니다. * **특징:** 가장 이상적이고 다루기 쉬운 케이스입니다. * **처리:** 데이터가 충분히 많다면, MCAR인 경우는 결측치가 있는 행을 삭제해도 편향이 발생하지 않습니다. (단, 데이터 손실은 감수해야 함)

3.2 2. MAR (Missing at Random, 임의 결측)

"결측 여부가 관측된 다른 변수와 관련이 있다"는 의미입니다.

결측이 발생한 변수(X_1) 자체와는 관련이 없지만, 우리가 관측할 수 있는 다른 변수(X_2)와는 관련이 있는 경우입니다.

* **예시:** 직장 내 괴롭힘에 대한 설문조사에서 '괴롭힘 경험' 문항에 결측치가 많습니다. 이 결측 여부가 '괴롭힘 경험' 자체와는 관련이 없을 수 있지만, '성별' 변수와는 관련이 있을 수 있습니다. (예: 남성이 여성보다 해당 문항 응답을 더 꺼림) * **특징:** '성별'이라는 관측된 변수를 활용하면 결측의 패턴을 설명할 수 있습니다. * **처리:** 모델링을 통해 결측치를 잘 처리할 수 있습니다. (예: 성별을 예측 변수로 사용하여 결측치 대체 모델 생성)

3.3 3. MNAR (Missing Not at Random, 비임의 결측)

"결측 여부가 누락된 값 자체 또는 관측되지 않은 변수와 관련이 있다"는 의미입니다.

* **예시 1 (누락된 값 자체):** '소득' 설문에서, 고소득자일수록 자신의 소득을 밝히기 꺼려 해서 응답을 하지 않는 경우. 즉, '소득'이라는 값 자체가 높을수록 결측이 될 확률이 높습니다. * **예시 2 (관측되지 않은 변수):** 임상시험에서 부작용이 심한 환자들이(관측되지 않은 '부작용' 변수) 시험을 중도 포기하여(결과값 결측) 데이터에서 누락되는 경우. * **특징:** 가장 다루기 어렵고 심각한 편향을 유발합니다. 우리가 그 이유를 알 수 없기 때문입니다. * **처리:** 통계적으로 완벽하게 해결하기 매우 어렵습니다.

주의사항

Q: 내 데이터가 어떤 유형인지 어떻게 알 수 있나요?
A: 알 수 없습니다.
안타깝게도 우리는 데이터만 보고 이 결측치가 MCAR, MAR, MNAR 중 무엇인지 통계적으로 완벽하게 증명할 수 없습니다. MNAR은 "관측되지 않은" 변수에 의해 발생할 수 있기 때문입니다.
따라서 데이터 분석가들은 **"우리의 데이터가 최소한 MAR이라고 가정하고, 관측된 다른 변수들을 최대한 활용하여 결측치를 모델링하자"**라는 실용적인 접근 방식을 취합니다.

특징	MCAR (완전 임의 결측)	MAR (임의 결측)	MNAR (비임의 결측)
결측 원인	완전 무작위	**관측된** 다른 변수	**누락된 값 자체** 또는 **숨겨진** 변수
예시	데이터 입력 실수	성별에 따라 소득 응답률 다름	고소득자가 소득 응답 거부
해결 난이도	쉬움 (삭제 가능)	중간 (모델링으로 해결 가능)	매우 어려움 (편향 피하기 힘듦)

Table 2: 결측 *3*유형 비교

4 결측치 처리(Imputation) 방법론

결측치를 단순히 삭제하거나 평균값으로 채우는 대신, 더 정교한 방법들을 사용해야 합니다.

4.1 처리 전 고려사항

1. **결측치가 어디에 있는가? (Y vs X)** * **예측 변수(X)에 결측:** 대부분의 처리 기법이 여기에 초점을 맞춥니다. * **반응 변수(Y)에 결측:** 매우 다루기 어렵습니다. Y값을 예측하는 것이 모델의 최종 목표인데, 그 Y값이 없기 때문입니다. 이 경우 해당 행을 삭제하는 것이 일반적입니다. 2. **변수 유형 (수치형 vs 범주형):** 처리 방법이 달라집니다. (예: 수치형은 k-NN, 범주형은 로지스틱 회귀) 3. **결측량 (Amount):** 만약 특정 변수가 60% 이상 결측치라면, 이 변수를 대체하는 것은 오히려 새로운 노이즈를 만드는 것일 수 있습니다. 이 경우 해당 변수(열)를 삭제하는 것을 고려할 수 있습니다.

4.2 방법 1: 결측치 표시자 변수 (Missingness Indicator)

이 방법은 결측치가 발생했다는 '사실 자체'가 중요한 정보를 담고 있을 수 있다고 가정합니다. (특히 MNAR의 경우에 유용)

* **아이디어:** "응답 거부"라는 제3의 그룹을 만듭니다. * **방법:** 1. 결측치가 있는 변수 X_1을 복사하여 두 개의 변수를 만듭니다. 2. X_1^* (대체 변수): X_1의 결측치를 0이나 평균 등 특정 값으로 모두 대체합니다. 3. $X_{1,miss}$ (표시자 변수): X_1에서 값이 누락되었으면 1, 아니면 0을 갖는 이진(binary) 변수를 만듭니다. 4. 모델을 학습시킬 때, 원래 변수인 X_1 대신 이 두 변수(X_1^*와 $X_{1,miss}$)를 함께 사용합니다.

□ **예제:**

예제 **예시: 결측치 표시자 변수 생성**
아래 표는 `X1`과 `X2`의 결측치를 0으로 대체하고, 결측 여부를 `X1_miss`, `X2_miss`로 표시한 예입니다. 이제 모델은 X_1^*와 $X_{1,miss}$를 보고, 'X1_miss가 1일 때(즉, X1이 누락되었을 때)'는 X_1^*의 0을 '관측된 0'과 다르게 취급할 수 있게 됩니다.

4.3 방법 2: 모델 기반 대체 (Model-based Imputation)

결측치를 '예측' 문제로 접근하는 방식입니다.

* **아이디어:** X_1 변수에 결측치가 있다면, 나머지 관측된 변수들($X_2, X_3, ...$)을 독립 변수로, X_1을 종속 변수로 하는 예측 모델을 만듭니다. * **절차:** 1. 데이터를 두 그룹으로 나눕니다. * **학습용(Train):** X_1이 관측된 모든 행. * **예측용(Test):** X_1이 누락된 모든 행. 2. 학습용 데이터로 $X_2, X_3, ... \rightarrow X_1$을 예측하는 모델(예: k-NN, 선형 회귀, 의사결정나무)을 학습시킵니다. 3. 학습된 모델을 예측용 데이터에 적용하여, 누락된 X_1 값을 예측하고 그 값으로 채워 넣습니다.

□ 예제:

예제 **예시: k-NN을 사용한 대체**
색깔(X, 관측됨)을 이용해 Y값(결측 존재)을 대체해봅시다. (k=2, 즉 가장 가까운 2개 사용)
[참고 이미지: *k-NN* 대체 시각화 - 색깔 기반으로 가장 가까운 *2*개 이웃의 평균값 사용] * **첫 번째 물음표 (?):** 색깔이 '중간 빨강'입니다. 가장 가까운 2개는 '진한 빨강'(Y=1)과 '밝은 빨강'(Y=0.5)입니다. * **대체:** 두 값의 평균인 $(1 + 0.5)/2 = 0.75$를 채워 넣습니다. * **두 번째 물음표 (?):** 색깔이 '노랑'입니다. 가장 가까운 2개는 '주황'(Y=0.1)과 '연두'(Y=10)입니다. * **대체:** 두 값의 평균인 $(0.1 + 10)/2 = 5.05$를 채워 넣습니다.

4.4 방법 3: 불확실성을 고려한 모델 기반 대체

위의 '모델 기반 대체'는 한 가지 큰 문제점을 가집니다. 바로 **"너무 완벽한"** 값을 채워 넣는다는 것입니다.

주의사항

결정론적 대체(Deterministic Imputation)의 함정
선형 회귀 모델로 결측치를 대체한다고 상상해봅시다. 예측된 값들은 모두 회귀선 위에 완벽하게 놓이게 됩니다. (아래 그림의 왼쪽)
하지만 실제 데이터는 어떻습니까? 항상 회귀선 주변에 흩어져 있습니다. (아래 그림의 오른쪽, 회색 점)
만약 우리가 모든 결측치를 회귀선 위의 완벽한 값으로만 채운다면, 데이터의 실제 '불확실성(분산)'이 사라지고 매우 인위적으로 좁은 분포를 갖게 됩니다. 이는 모델이 현실을 과도하게 확신하게 만듭니다.

[참고 이미지: 결정론적 *vs.* 확률적 예측 시각화 - 확률 모델은 예측에 불확실성을 반영]

해결책: 예측에 무작위성(불확실성)을 더하자!

* **k-NN 대체 시:** k개의 이웃을 찾은 뒤, 그 값들을 '평균'내는 대신 k개 중 하나를 **무작위로 샘플링**하여 채워 넣습니다. * (위의 '노랑' 예시: 5.05를 채우는 대신, 50% 확률로 0.1, 50% 확률로 10을 뽑아 넣습니다.) * **선형 회귀 대체 시:** 예측값 $\hat{y}$을 구한 뒤, 학습 데이터의 실제 잔차(residual, ϵ) 중 하나를 **무작위로 샘플링**하여 $\hat{y} + \epsilon$을 채워 넣습니다. * **의사결정나무 대체 시:** 예측값이 속한 최종 노드(leaf)에 여러 개의 학습 데이터가 있다면, 그 값들을 '평균'내는 대신 그중 하나를 **무작위로 샘플링**하여 채워 넣습니다.

이러한 '불확실성을 고려한 대체'는 데이터의 원래 분포와 분산을 보존하는 데 훨씬 효과적입니다.

4.5 여러 변수에 결측치가 있는 경우 (Iterative Imputation)

만약 X_1, X_2, X_3 모두에 결측치가 있다면 어떻게 해야 할까요? 이는 "닭과 달걀의 문제"와 같습니다. X_1을 예측하려면 X_2, X_3가 필요한데, X_2, X_3에도 결측치가 있습니다.

해결책: 반복적(Iterative)으로 예측합니다.

1. **초기화:** X_1, X_2, X_3의 모든 결측치를 일단 '평균값'으로 채웁니다. 2. **1라운드 (X1 예측):** (평균으로 채워진) X_2, X_3를 이용해 X_1의 결측치를 예측하고 업데이트합니다. 3. **1라운드 (X2 예측):** (방금 업데이트된) X_1과 (평균으로 채워진) X_3를 이용해 X_2의 결측치를 예측하고 업데이트합니다. 4.

1라운드 (X3 예측): (업데이트된) X_1, X_2를 이용해 X_3의 결측치를 예측하고 업데이트합니다. 5. **2라운드 이후:** 1 4의 과정을 X_1, X_2, X_3의 대체값들이 더 이상 크게 변하지 않을 때까지 (수렴, converge) 여러 번 반복합니다.

4.6 Sklearn을 사용한 구현

이러한 복잡한 과정들은 `sklearn.impute` 모듈에 구현되어 있습니다.

```
from sklearn.impute import SimpleImputer
from sklearn.impute import IterativeImputer
from sklearn.impute import KNNImputer
from sklearn.impute import MissingIndicator

# 1. 단순대체평균 (, 중앙값, 최빈값등 )
imputer_simple = SimpleImputer(strategy='mean')

# 2. 결측치표시자
indicator = MissingIndicator()

# 3. k-NN 기반대체모델 ( 기반)
imputer_knn = KNNImputer(n_neighbors=5)

# 4. 반복적대체가장 ( 정교한방법 )
# 다른모든피처를사용하여각피처의결측치를예측
imputer_iterative = IterativeImputer(max_iter=10, random_state=0)
```

Listing 1: sklearn의 주요 Imputer

5 블랙박스 모델 해석 및 시각화

5.1 시각화의 목표

데이터 시각화는 여러 목표를 갖습니다.

* (모델링 전) 데이터 탐색 및 가설 수립 (EDA) * (모델링 후) **모델 결과 커뮤니케이션**

모델이 복잡해질수록 "모델이 왜 이런 예측을 했는가?"를 설명하기 어려워집니다. 이 섹션은 모델링 후의 커뮤니케이션에 초점을 맞춥니다.

5.2 왜 모델 해석이 필요한가? (Parametric vs. Non-parametric)

1. **파라메트릭 모델 (Parametric Models):** * 예: 선형 회귀, 로지스틱 회귀 * 모델이 $Y = \beta_0 + \beta_1 X_1 + ...$ 처럼 간단한 수식으로 정의됩니다. * 우리는 계수(coefficient) β_1을 보고 "X1이 1단위 증가할 때 Y는 β_1만큼 증가(또는 감소)한다"라고 명확하게 해석할 수 있습니다. (화이트박스 모델) 2. **비파라메트릭 모델 (Non-parametric Models):** * 예: k-NN, 의사결정나무, 랜덤 포레스트 * 모델이 복잡한 규칙(Tree)이나 거리 계산(k-NN)의 조합으로 이루어집니다. * "β_1"처럼 해석할 수 있는 간단한 계수가 없습니다. * 이처럼 내부 작동 원리를 직관적으로 파악하기 어려운 모델을 **블랙박스(Black Box)** 모델이라고 부릅니다.

블랙박스 모델이라도, 우리는 모델이 '어떻게' 예측하는지 이해해야 합니다. 이때 사용하는 기법이 **부분 의존성 플롯(PDP)**입니다.

5.3 부분 의존성 플롯 (Partial Dependence Plots, PDP)

PDP는 "다른 모든 변수들을 평균(또는 특정 값)으로 고정했을 때, 내가 관심 있는 변수 하나(X_1)를 변화시키면 모델의 예측값이 어떻게 변하는가?"를 보여주는 그래프입니다.

□ 예제:

예제 **PDP의 직관적 비유: 오디오 믹서**
PDP는 오디오 믹서(Mixer)와 같습니다. 훌륭한 음악(모델 예측)은 보컬, 드럼, 베이스(여러 변수)가 조합된 결과입니다.
PDP는 이 음악에서 **'베이스(X1)'가 전체 사운드에 어떤 영향을 주는지** 알고 싶을 때, '보컬'과 '드럼'(다른 변수)의 볼륨을 '중간'으로 고정해 놓고, '베이스'의 볼륨만 0에서 100까지 싹 돌려보면서 소리의 변화(Y 예측값)를 녹음하는 것과 같습니다.

—

5.3.1 예제 1: 단일 변수 모델 시각화

먼저 간단한 모델로 시작합니다. 심장병(AHD, 0 or 1)을 최대 심박수(MaxHR) 하나만으로 예측하는 k-NN(k=50) 모델을 만들었습니다.

* **모델:** `AHD ~ MaxHR` * **시각화:** 'MaxHR' 값을 70부터 200까지 촘촘하게 만들고(synthetic X), 각 값에 대해 모델이 예측하는 '심장병 확률'을 계산하여 점을 찍어 연결합니다.

[참고 이미지: *PDP 시각화 - MaxHR vs.* 심장병 확률의 부정적 관계]

* **해석:** 이 그래프(PDP)는 'MaxHR'이 낮을수록 심장병 확률이 70% 이상으로 높고, 'MaxHR'이 높을수록(즉, 건강할수록) 확률이 20% 근처로 낮아지는 **부정적(negative) 관계**를 모델이 학습했음을 보여줍니다.

—

5.3.2 예제 2: 다중 변수 모델의 PDP (상호작용이 없는 경우)

이제 더 복잡한 모델을 만듭니다. 'MaxHR'뿐만 아니라 'Age', 'Sex', 'RestBP' 등 10개의 변수를 사용했습니다.

* **모델:** `AHD  MaxHR + Age + Sex + ...` * **시각화 (PDP):** 'MaxHR'의 영향을 보기 위해, 다른 9개 변수('Age', 'Sex' 등)를 모두 **"전체 데이터의 중앙값(median)"**으로 고정합니다. (즉, '평균적인 환자'를 가정) * 그런 다음, 이 '평균적인 환자'의 'MaxHR'만 70에서 200으로 바꾸면서 심장병 확률을 예측합니다.

[참고 이미지: 다중 변수 *PDP* - 평균 환자에 대한 *MaxHR* 영향력 시각화]

* **해석:** 여러 변수를 추가했지만, '평균적인 환자'에 대한 'MaxHR'의 영향력은 예제 1과 거의 유사한 부정적 관계를 보입니다.

—

5.3.3 예제 3: PDP를 이용한 상호작용(Interaction) 발견

PDP의 진정한 힘은 '평균적인 환자'가 아닌 '특정 그룹'에 대한 예측을 비교할 때 나옵니다.

* **가설:** 'MaxHR'이 심장병에 미치는 영향은 'Age' (나이)에 따라 다르지 않을까? * **시각화 (PDP 비교):** 3개의 PDP를 한꺼번에 그립니다. 1. (주황색) 'Age' = **중앙값** (약 55세), 다른 변수도 중앙값 2. (초록색) 'Age' = **최대값** (약 77세), 다른 변수는 중앙값 3. (빨간색) 'Age' = **최소값** (약 29세), 다른 변수는 중앙값

[참고 이미지: 상호작용 *PDP* - 다양한 *Age* 값에 따른 *MaxHR* 효과 변화]

* **해석:** * 그래프가 세 그룹(다른 나이)에 대해 다르게 그려집니다! 이는 모델이 **'MaxHR'와 'Age' 간의 상호작용**을 학습했다는 뜻입니다. * 'MaxHR'이 낮을 때는 (왼쪽, < 110) 나이와 상관없이 모두 확률이 높습니다. * 'MaxHR'이 높을 때는 (오른쪽, > 160) 나이의 영향이 커집니다. 젊은 환자(빨간색, 최소값)는 확률이 15%까지 떨어지지만, 나이 많은 환자(초록색, 최대값)는 확률이 25% 정도로 상대적으로 높게 유지됩니다. * **결론:** 이 모델에 따르면, 최대 심박수가 높은 것(운동)은 젊은 사람에게 더 큰 심장병 예방 효과가 있습니다.

6 효과적인 시각화의 원칙 (부록)

6.1 나쁜 시각화의 예

모델 해석뿐만 아니라 모든 데이터 시각화에서 '나쁜 시각화'는 의미를 왜곡합니다. [참고 이미지: 잘못된 시각화 예시 - 하버드 *GPA* 인플레이션을 왜곡한 차트]

위 차트는 하버드의 GPA 인플레이션을 보여주려 했으나, 최악의 시각화 중 하나입니다.

* **문제점:** 2005년부터 2017년까지의 '3.67'과 2017년 이후의 '4.00'은 완전히 다른 척도(하나는 숫자, 하나는 등급)를 의미할 수 있으며, 막대그래프로 수치를 표현했지만 모든 막대의 높이가 동일하여 시각적 정보를 전혀 주지 못합니다. 이는 데이터를 조작하고 청중을 속이는 행위입니다.

6.2 좋은 시각화의 원칙

역사적으로 가장 위대한 데이터 시각화들은 다음과 같은 원칙을 따릅니다. (예: 존 스노우의 콜레라 지도, 나이팅게일의 로즈 차트, 미나르의 나폴레옹 행군도)

1. **그래픽 무결성 (Graphical Integrity):** 데이터를 왜곡하거나 속이지 않아야 합니다. (예: Y축을 0에서 시작하기, 척도 통일하기) 2. **단순함 (Keep it simple):** 불필요한 장식(3D 효과, 그림자, 과도한 색상)을 제거하여 '데이터 잉크 비율'을 높여야 합니다. 3. **올바른 디스플레이 사용 (Use the right display):** * (Good) **위치(Position), 길이(Length):** 사람이 가장 정확하게 인지. (예: 막대 차트, 산점도) * (Bad) **각도(Angle), 면적(Area):** 사람이 크기를 과소/과대평가하기 쉬움. (예: 파이 차트, 도넛 차트) 4. **전략적인 색상 사용 (Use color strategically):** * **범주형(Qualitative):** 서로 구분이 명확한 색상 사용 (예: 정당, 과일 종류) * **순차형(Sequential):** 하나의 색상을 연한 색 진한색으로 표현 (예: 인구 밀도 $0 \rightarrow 100$) $***$ $(Diverging)$: $**$ (0) (: $-100 \rightarrow 0 \rightarrow +100$) $*$ / .5. $**$ $(Knowyouraudience)$: $**$ .6. $**$ $(Tellastory)$: $**$ $'-\ -'$. $(Annotations, CallOutBoxes)$.

7 빠르게 훑어보기 (1-Page Summary)

1. 결측치(Missingness)란?

데이터에 값이 누락된 '구멍' (`NaN`). 그냥 삭제하거나 평균값으로 채우면 **편향(Bias)**이 발생하여 모델이 현실을 잘못 예측하게 됨.

2. 결측의 3유형 (원인)

- **MCAR (완전 임의):** 완전 무작위 (예: 오타). 삭제해도 편향 없음 (데이터는 손실됨).
- **MAR (임의):** 관측된 변수(예: 성별)와 관련됨. 모델링으로 해결 가능.
- **MNAR (비임의):** 값 자체(예: 고소득) 또는 숨겨진 변수(예: 부작용)와 관련됨. 해결 매우 어려움.

현실: 어떤 유형인지 증명 불가. 보통 MAR로 가정하고 모델링.

3. 결측치 처리(Imputation) 전략

- **표시자 변수 (Indicator):** "결측됨" 자체를 정보로 활용. (결측=1, 아닌=0인 새 변수 추가)
- **모델 기반 대체 (k-NN, Regression):** 다른 변수들로 결측치를 '예측'하여 채움.
- **불확실성 고려 대체 (Best):** 예측값($\hat{y}$)에 무작위성(예: 잔차 ϵ)을 더해 $\hat{y} + \epsilon$로 채움. (데이터의 실제 분포를 보존하기 위해)
- **반복적 대체 (Iterative):** 여러 변수에 결측치가 있을 때, 수렴할 때까지 서로를 예측하며 반복적으로 채움 (`IterativeImputer`).

4. 블랙박스 모델 해석 (PDP)

PDP (부분 의존성 플롯): 복잡한 모델(k-NN, 트리 등)을 해석하는 도구.
"다른 변수들은 '평균'으로 고정하고, 관심 변수 X 하나만 바꿨을 때 모델의 예측 Y가 어떻게 변하는지" 보여주는 그래프.
만약 '나이(A) 그룹'별로 그린 PDP와 '나이(B) 그룹'별로 그린 PDP의 모양이 다르다면, 모델이 '나이'와 '관심 변수 X' 간의 **상호작용(Interaction)**을 학습했다는 의미.