CASE STUDY 1 · CHAPTER 2 · PART 1

조류 대발생 예측 (1)
데이터 탐색과 결측치 처리

Predicting Algae Blooms — 실제 유럽 하천 수질 데이터로 배우는 R 데이터마이닝 파이프라인

원저: Luis Torgo, 'Data Mining with R' (2003) Chapter 2 를 바탕으로, Analysis.txt / Eval.txt / Sols.txt 데이터를 R로 직접 실행한 결과로 재구성

OVERVIEW

학습 목표

1. **문제와 데이터 이해**
   조류 대발생 예측 문제의 배경과 Analysis/Eval/Sols 데이터 구조를 설명할 수 있다.

2. **탐색적 데이터 분석**
   히스토그램·박스플롯·조건부 그래프로 변수의 분포와 이상치를 파악할 수 있다.

3. **결측치 처리**
   제거·평균/중앙값 대체·상관관계·유사사례 기반 등 결측치 처리 전략을 이해하고 R로 구현할 수 있다.

4. **다음 시간 준비**
   모델링(Part 2)에 사용할 완전한(clean) 데이터셋을 완성한다.

PROBLEM

문제 설명: 왜 조류 대발생을 예측할까?

- 하천의 유해 조류(algae) 대량 발생은 생태계와 수질에 심각한 영향을 미치는 문제
- 유럽 여러 하천에서 약 1년에 걸쳐 수질 샘플을 채취, 화학적 특성과 7종 유해 조류의 발생 빈도를 기록
- 화학적 성분 측정은 저렴하고 자동화하기 쉬운 반면, 조류를 현미경으로 식별하는 생물학적 분석은 비싸고 느림
- 목표: 화학적 특성만으로 조류 빈도를 예측하는 모델을 만들어, 저렴하고 자동화된 조기 경보 시스템을 구축

> *화학 모니터링은 자동화가 쉽고 저렴하지만, 실제 조류 식별(생물학적 분석)은 숙련된 인력과 시간이 필요합니다. 이 간극을 메우는 것이 예측 모델의 역할입니다.*

DATA

데이터 설명

파일	샘플 수	내용
Analysis.txt	200개	학습(훈련)용 수질 샘플 — 설명변수 11개 + 조류 빈도 7개
Eval.txt	140개	평가(테스트)용 수질 샘플 — 설명변수 11개만 포함 (조류 빈도 없음)
Sols.txt	140개	Eval.txt 샘플들의 실제 조류 빈도 정답 — 모델 성능 검증용

- 설명변수 11개 = 계절/강 크기/유속(범주형 3개) + 화학 성분 8개(mxPH, mnO2, Cl, NO3, NH4, oPO4, PO4, Chla)
- 목표변수 7개(a1~a7) = 7종 유해 조류 각각의 발생 빈도 (알려지지 않은 실제 조류 이름)
- 결측값은 텍스트 파일에서 문자열 'XXXXXXX'로 표시됨

LOADING DATA

R로 데이터 불러오기

R CONSOLE

```
> algae <- read.table('Analysis.txt', header=F, dec='.',
+   col.names=c('season','size','speed','mxPH','mnO2','Cl',
+   'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',
+   'a5','a6','a7'),
+   na.strings=c('XXXXXXX'))
```

- header=F : 파일 첫 줄에 변수명이 없음 · col.names : 각 열에 붙일 이름을 직접 지정
- na.strings='XXXXXXX' : 이 문자열을 결측값(NA)으로 해석

실행 결과: nrow(algae) = 200, ncol(algae) = 18 → 200개 샘플 × 18개 변수의 데이터프레임이 만들어졌습니다.

EXPLORATION

기술통계 요약: summary(algae)

변수	요약
season / size / speed	autumn 40 · spring 53 · summer 45 · winter 62 / large 45 · medium 84 · small 71 / high 84 · low 33 · medium 83
mxPH	평균 8.01, 중앙값 8.06, 범위 [5.6, 9.7], 결측 1개
Cl / NH4	Cl 결측 10개(가장 많음) · NH4 평균 501.3 ≫ 중앙값 103.2 → 매우 치우친(skewed) 분포
Chla / a1	Chla 결측 12개 · a1(조류1)은 평균 16.9, 최댓값 89.8로 변동이 큼

summary()는 범주형 변수는 각 값의 빈도를, 수치형 변수는 최소/1사분위/중앙값/평균/3사분위/최대값과 결측치 개수를 보여줍니다.

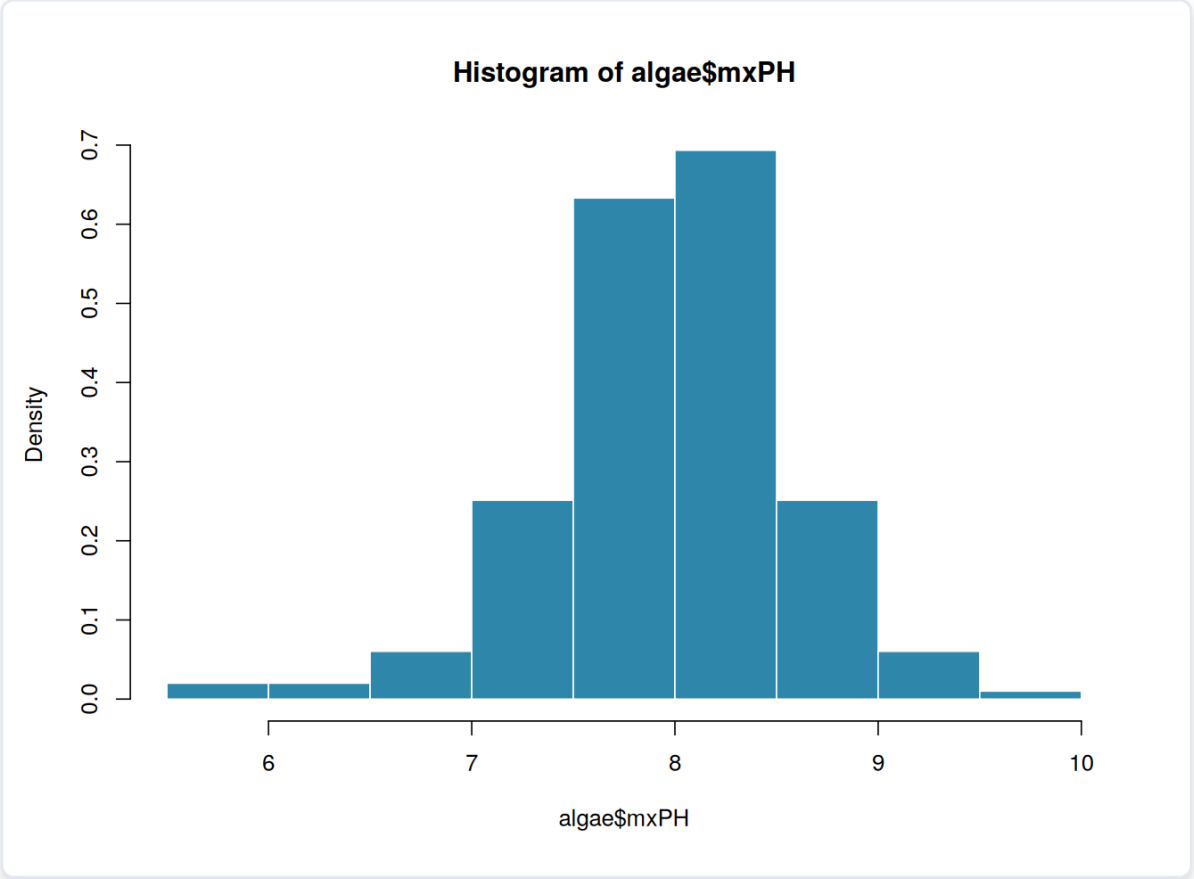
VISUALIZATION

히스토그램으로 분포 살펴보기

R CONSOLE

```
> hist(algae$mxPH, prob=T)
```

- prob=T : 도수 대신 확률(밀도)로 표시
- mxPH는 평균 주변에 값이 고르게 모여있는, 정규분포에 가까운 형태



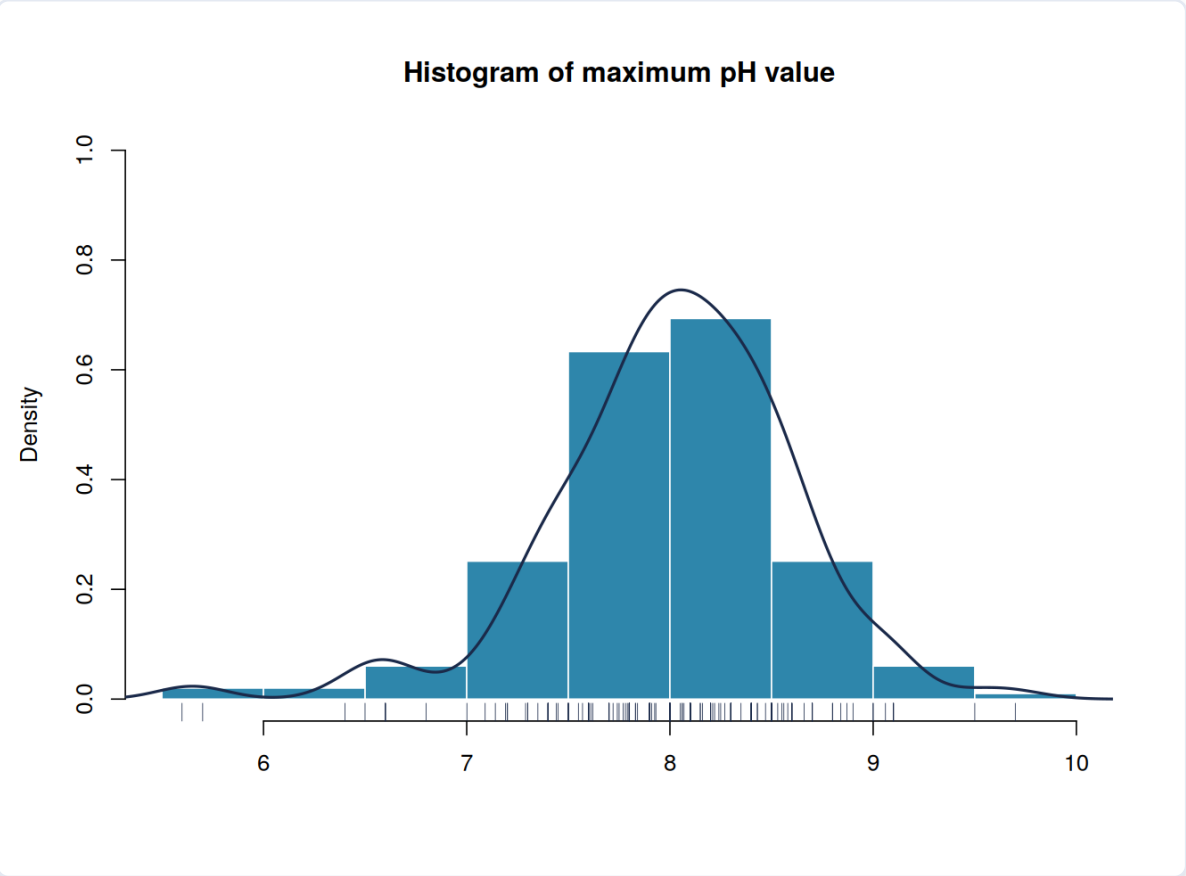
VISUALIZATION

밀도곡선과 러그(rug)를 더한 히스토그램

R CONSOLE

```
> hist(algae$mxPH, prob=T, xlab='',
+   main='Histogram of maximum pH value',
ylim=0:1)
> lines(density(algae$mxPH, na.rm=T))
> rug(jitter(algae$mxPH))
```

- density() : 부드러운 커널 밀도 추정 곡선
- rug() : 실제 관측값을 x축 근처에 눈금으로 표시 → 이상치 발견에 유용
- 그래프 하단에서 다른 값들과 뚝 떨어진 낮은 값 2개를 확인할 수 있음



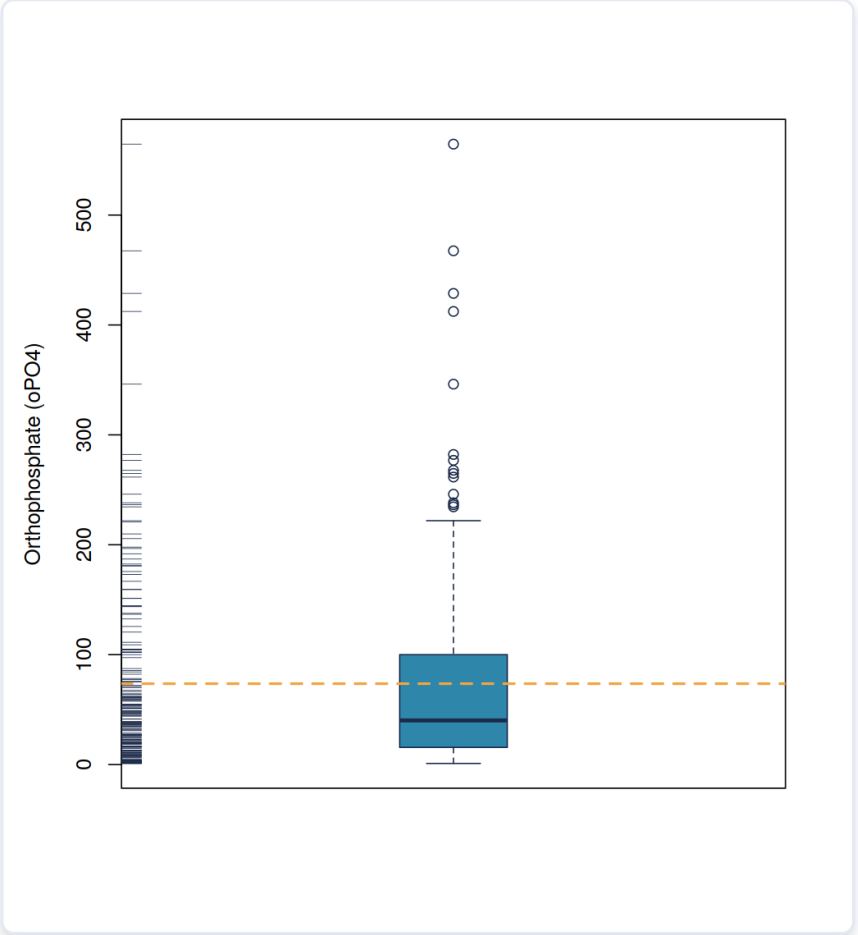
VISUALIZATION

박스플롯(Box Plot)으로 이상치 확인

R CONSOLE

```
> boxplot(algae$oPO4, boxwex=0.15,
+   ylab='Orthophosphate (oPO4)')
> rug(jitter(algae$oPO4), side=2)
> abline(h=mean(algae$oPO4, na.rm=T), lty=2)
```

- 상자 = 1·3사분위, 상자 안 굵은 선 = 중앙값
- 위·아래 수염 밖 원 = 이상치(outlier) 후보
- 점선 = 평균값 — 중앙값보다 훨씬 높음 → 오른쪽으로 치우친(skewed) 분포임을 시사



VISUALIZATION

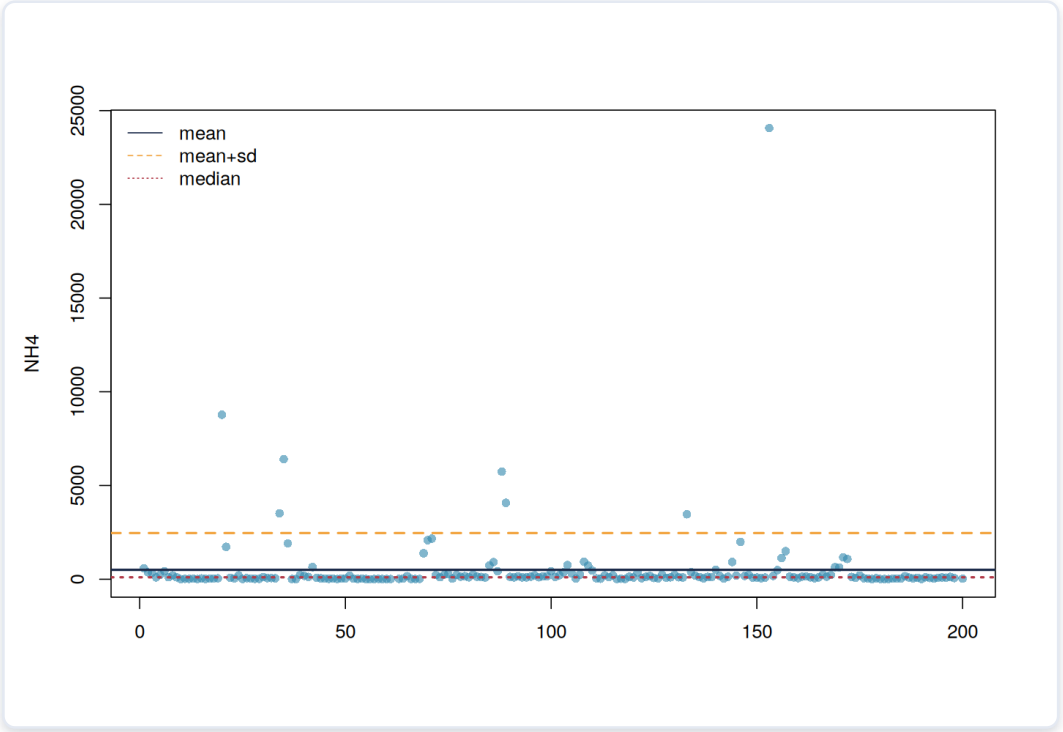
이상치가 있는 샘플 찾아내기

R CONSOLE

```
> plot(algae$NH4, xlab='')
> abline(h=mean(algae$NH4,na.rm=T), lty=1)
> abline(h=mean(algae$NH4,na.rm=T)+
+           sd(algae$NH4,na.rm=T), lty=2)
> algae[!is.na(algae$NH4) & algae$NH4>19000,]
```

실행 결과 (153번 샘플): NH4 = 24064 — 다른 관측치들과 확연히 동떨어진 값. season=autumn, size=medium, speed=high.

- 그래프로 이상치를 발견한 뒤, 논리 인덱싱으로 해당 샘플의 행 전체를 확인할 수 있음
- identify() 함수를 쓰면 그래프에서 마우스로 클릭해 대화형으로 점을 찾을 수도 있음

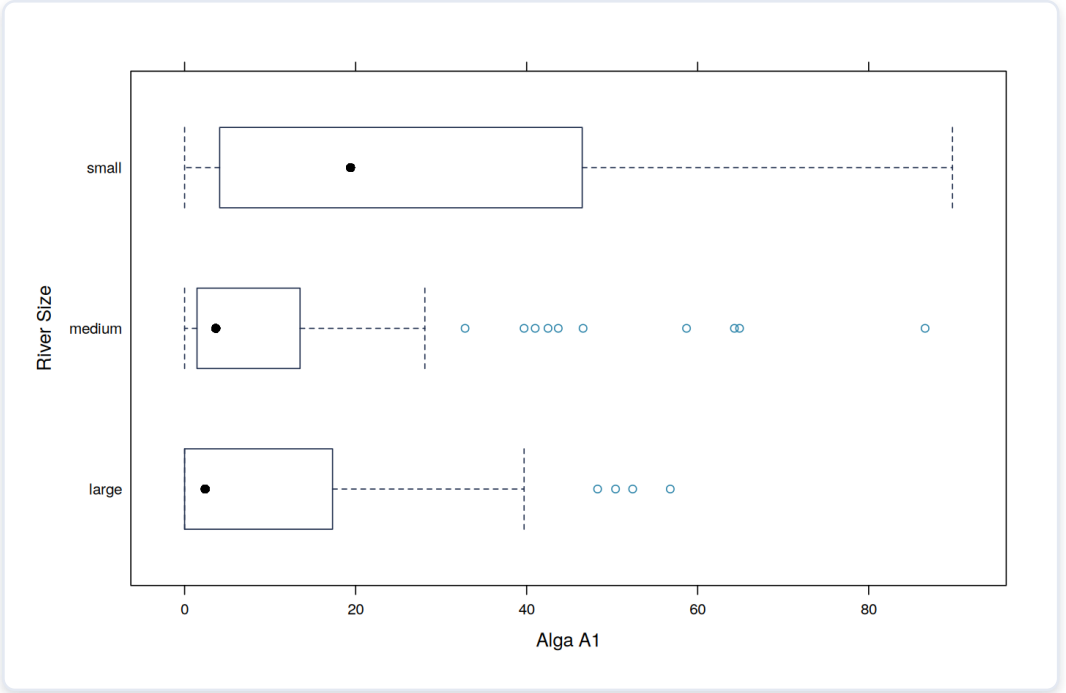


조건부 박스플롯: 강 크기별 조류 a1 분포

R CONSOLE

```
> library(lattice)
> bwplot(size ~ a1, data=algae,
+   ylab='River Size', xlab='Alga A1')
```

- lattice 패키지의 조건부(conditioned) 그래프: 범주형 변수별로 그래프를 나누어 비교
- 작은 강(small)에서 조류 a1의 빈도가 상대적으로 더 높게 나타나는 경향을 관찰할 수 있음



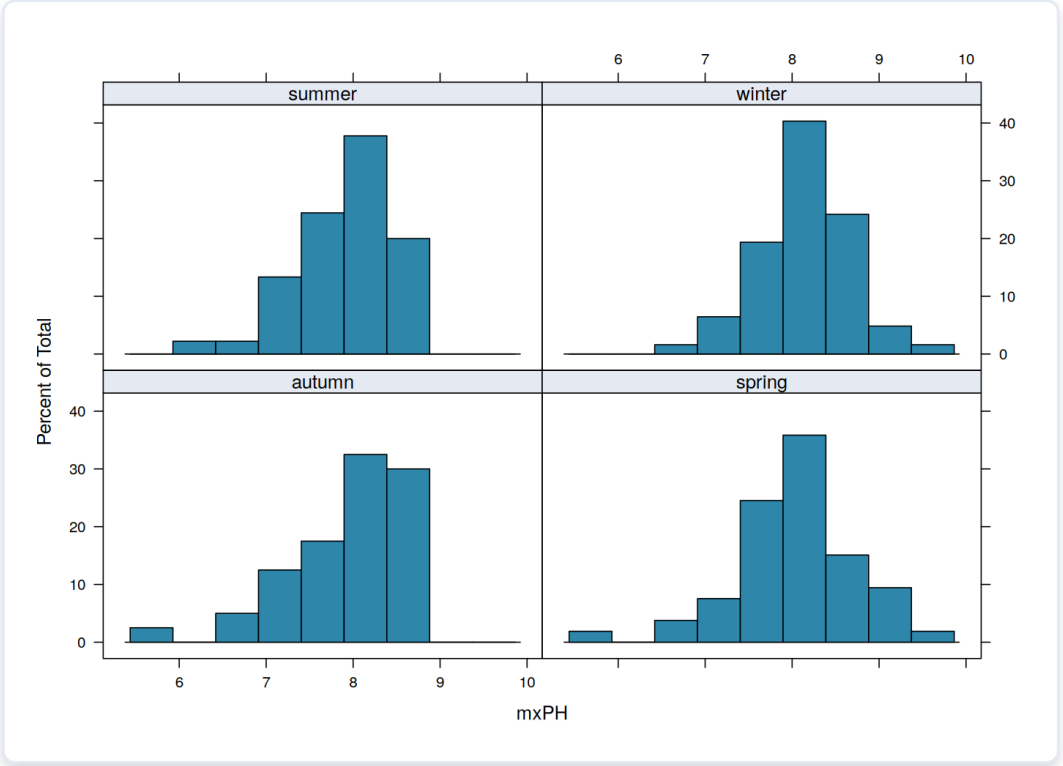
VISUALIZATION · LATTICE

조건부 히스토그램: 계절별 mxPH 분포

R CONSOLE

```
> histogram(~ mxPH | season, data=algae)
```

- '~ mxPH | season' : season의 각 값(봄/여름/가을/겨울)마다 별도의 히스토그램을 그림
- 네 계절의 히스토그램 모양이 서로 비슷함 → mxPH는 계절의 영향을 크게 받지 않는 것으로 보임
- 같은 방식으로 size, speed 등 다른 범주형 변수와의 관계도 탐색 가능



MISSING VALUES

결측치가 있는 샘플 확인하기

R CONSOLE

```
> algae[!complete.cases(algae),]
> nrow(algae[!complete.cases(algae),])
[1] 16
```

실행 결과: 결측치가 하나라도 있는 샘플은 16개 (28, 38, 48, 55~63, 116, 161, 184, 199번). 이 중 62번과 199번은 11개 설명변수 중 6개가 결측 — 거의 정보가 없는 샘플.

- complete.cases() : 결측값이 하나도 없는 '완전한' 행이면 TRUE
- '!' 연산자로 논리값을 반대로 뒤집어 결측치가 있는 행만 선택

> *선형회귀처럼 일부 모델은 결측값이 있으면 아예 학습이 불가능합니다. 실제 데이터에서는 결측치가 매우 흔하므로, 이를 다루는 전략이 필수적입니다.*

MISSING VALUES

결측치 처리 전략 4가지

1

① 제거하기

결측치가 있는 샘플(행)을 아예 데이터셋에서 제거. 결측 비율이 낮을 때 간단하고 효과적

2

② 최빈값으로 채우기

평균 또는 중앙값 등 대표값으로 결측치를 채움. 빠르지만 편향(bias)이 생길 수 있음

3

③ 상관관계 탐색으로 채우기

다른 변수와의 상관관계(회귀식)를 이용해 더 그럴듯한 값을 추정

4

④ 유사 사례 탐색으로 채우기

가장 비슷한 다른 샘플들을 찾아 그 값들로 결측치를 추정 (최근접 이웃)

STRATEGY 1 · REMOVAL

전략 ①: 결측치가 있는 샘플 제거하기

R CONSOLE

```
> algae <- na.omit(algae)
```

- na.omit() : 결측값이 하나라도 있는 행을 모두 제거
- 결측 비율이 작을 때(16/200 ≈ 8%) 합리적인 선택이 될 수 있음
- 62번, 199번 샘플처럼 결측이 유난히 많은(11개 중 6개) 샘플은 채우기보다 제거하는 것이 안전

R CONSOLE · 일부만 제거

```
> algae <- algae[-c(62,199),]
```

이후 모델링(Part 2)에서는 이 두 샘플만 제거한 198개 데이터를 기본으로 사용합니다.

STRATEGY 2 · CENTRAL VALUE

전략 ②: 평균 또는 중앙값으로 채우기

R CONSOLE

```
> algae[48,'mxPH'] <- mean(algae$mxPH, na.rm=T)
> algae[is.na(algae$Chla),'Chla'] <-
+   median(algae$Chla, na.rm=T)
```

실행 결과: mean(mxPH, na.rm=T) = 8.0117 | mean(Chla, na.rm=T) = 13.971, median(Chla, na.rm=T) = 5.475

- mxPH는 정규분포에 가까우므로 평균으로 채움이 합리적
- Chla는 오른쪽으로 치우친(skewed) 분포 — 평균(13.971)이 대표값으로 부적절하므로 중앙값(5.475)을 사용
- 장점: 빠르고 구현이 쉬움 / 단점: 변수의 분산을 인위적으로 줄이고 편향을 일으킬 수 있어 대용량 데이터가 아니면 권장되지 않음

STRATEGY 3 · CORRELATION

전략 ③: 변수 간 상관관계 탐색

R CONSOLE

```
> symnum(cor(algae[,4:18], use="complete.obs"))
```

- symnum() : 상관행렬을 기호(공백/./,/*)로 요약해 한눈에 보기 쉽게 표시
- 대부분의 변수 쌍은 상관관계가 거의 없음(공백)
- 예외 두 가지: NH4-NO3 (약한 상관, ',') 그리고 PO4-oPO4 (강한 상관, '*' → 0.9 이상)
- PO4는 oPO4(용존 인산염)를 포함하는 총 인산염이므로 강한 상관은 화학적으로도 자연스러움

실행 결과 (일부)

```
      mP mO Cl NO NH o P Ch
mxPH  1
mnO2     1
Cl          1
NO3            1
NH4            , 1
oPO4     .  .         1
PO4      .  .         * 1
Chla  .                     1
```

STRATEGY 3 · CORRELATION

상관관계로 발견한 회귀식으로 채우기

R CONSOLE

```
> lm(oPO4 ~ PO4, data=algae)
Coefficients:
(Intercept)          PO4
   -15.6142       0.6466
```

oPO4 = -15.6142 + 0.6466 × PO4 (실제 실행 결과, 책의 수치와 정확히 일치)

R CONSOLE

```
> fillPO4 <- function(oP) {
+   if (is.na(oP)) return(NA)
+   else return((oP+15.6142)/0.6466)
+ }
> algae[is.na(algae$PO4),'PO4'] <-
+   sapply(algae[is.na(algae$PO4),'oPO4'], fillPO4)
```

- 직접 발견한 선형관계를 함수로 만들어 sapply()로 모든 결측치에 일괄 적용
- 실행 결과: fillPO4(6.5) = 34.2 — 하나의 oPO4 값에서 대응하는 PO4 예측값을 얻음
- 이 방법은 강한 상관관계가 있는 변수 쌍에서만 신뢰할 수 있음

STRATEGY 4 · SIMILARITY

전략 ④: 가장 비슷한 사례로 채우기

R CONSOLE

```
> library(cluster)
> dist.mtx <- as.matrix(daisy(algae, stand=T))
> sort(dist.mtx['38',])[1:10]
```

실행 결과 — 38번 샘플과 가장 가까운 이웃 10개(자기 자신 제외): 54, 22, 64, 11, 30, 25, 53, 37, 24, 18
38번 샘플의 mnO2가 결측 → 이웃들의 mnO2 중앙값으로 채움: median = 10

- daisy() : 범주형·수치형이 섞인 데이터에서도 거리(유클리드 거리 기반)를 계산 (cluster 패키지)
- stand=T : 변수마다 척도가 다르므로 평균 0, 표준편차 1로 표준화한 뒤 거리 계산
- 가장 가까운 10개 이웃의 중앙값으로 결측치를 대체

STRATEGY 4 · SIMILARITY

central.value() 함수로 전체 자동화하기

R CONSOLE

```
> central.value <- function(x) {
+   if (is.numeric(x)) median(x, na.rm=T)
+   else if (is.factor(x)) levels(x)[which.max(table(x))]
+ }
> for(r in which(!complete.cases(algae)))
+   algae[r, which(is.na(algae[r,]))] <-
+     apply(data.frame(algae[nbrs(r), which(is.na(algae[r,]))]),
+           2, central.value)
```

- central.value() : 수치형이면 중앙값, 범주형(factor)이면 최빈값을 반환하는 범용 함수
- for문으로 결측치가 있는 모든 샘플에 대해 자동으로 이웃 기반 채우기를 수행
- 실행 결과: sum(!complete.cases(clean.algae)) = 0 → 198개 샘플 모두 결측치 없이 완성!

SUMMARY

Part 1 정리

- 조류 대발생 예측 문제: 화학 성분으로 7종 조류의 빈도를 예측하는 회귀 문제
- 히스토그램·박스플롯·조건부(lattice) 그래프로 분포·이상치·변수 간 관계를 탐색
- 결측치 처리 4가지 전략: 제거, 평균/중앙값 대체, 상관관계 활용, 유사 사례(최근접 이웃) 활용
- 유사 사례 기반 방법으로 완성한 clean.algae (198개 샘플, 결측치 0개) — Part 2 모델링에 사용

다음 시간 (Part 2)

선형회귀·회귀나무 모델 만들기, 교차검증으로 성능 평가하기, 140개 테스트 샘플의 조류 빈도 최종 예측하기

PRACTICE

직접 해보기

1

다른 변수 시각화

hist()와 boxplot()을 이용해 Cl 또는 NO3 변수의 분포를 그려보고, 치우친 정도를 관찰해보세요.

2

결측치 세어보기

table(is.na(algae$Cl))로 Cl 변수의 결측치 개수를 직접 확인해보세요.

3

채우기 전략 비교

mxPH 48번 샘플을 평균으로 채우는 방법과, 유사 사례(이웃)로 채우는 방법의 결과값을 비교해보세요.