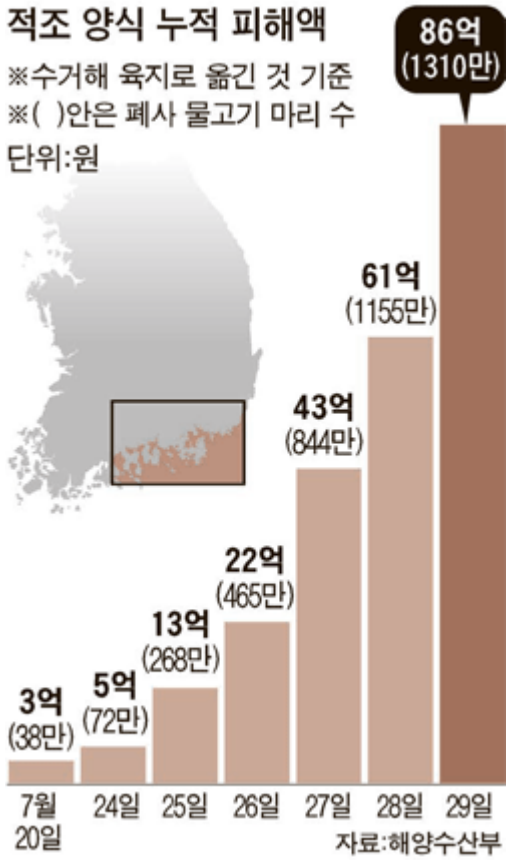
RED TIDE CLASSIFICATION

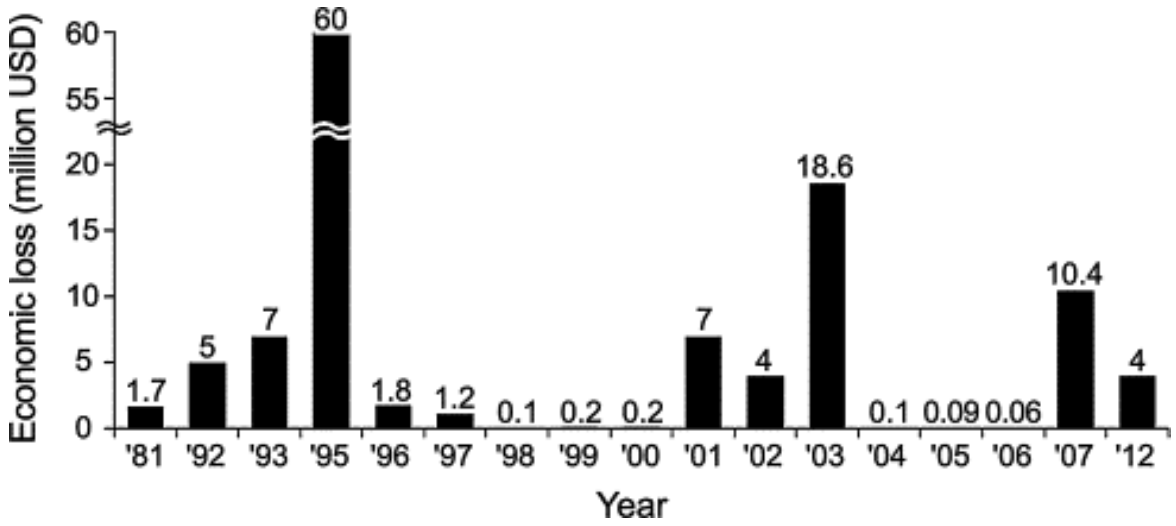
ITWILL 한수정

CONTENTS

1 Background

About Red Tide

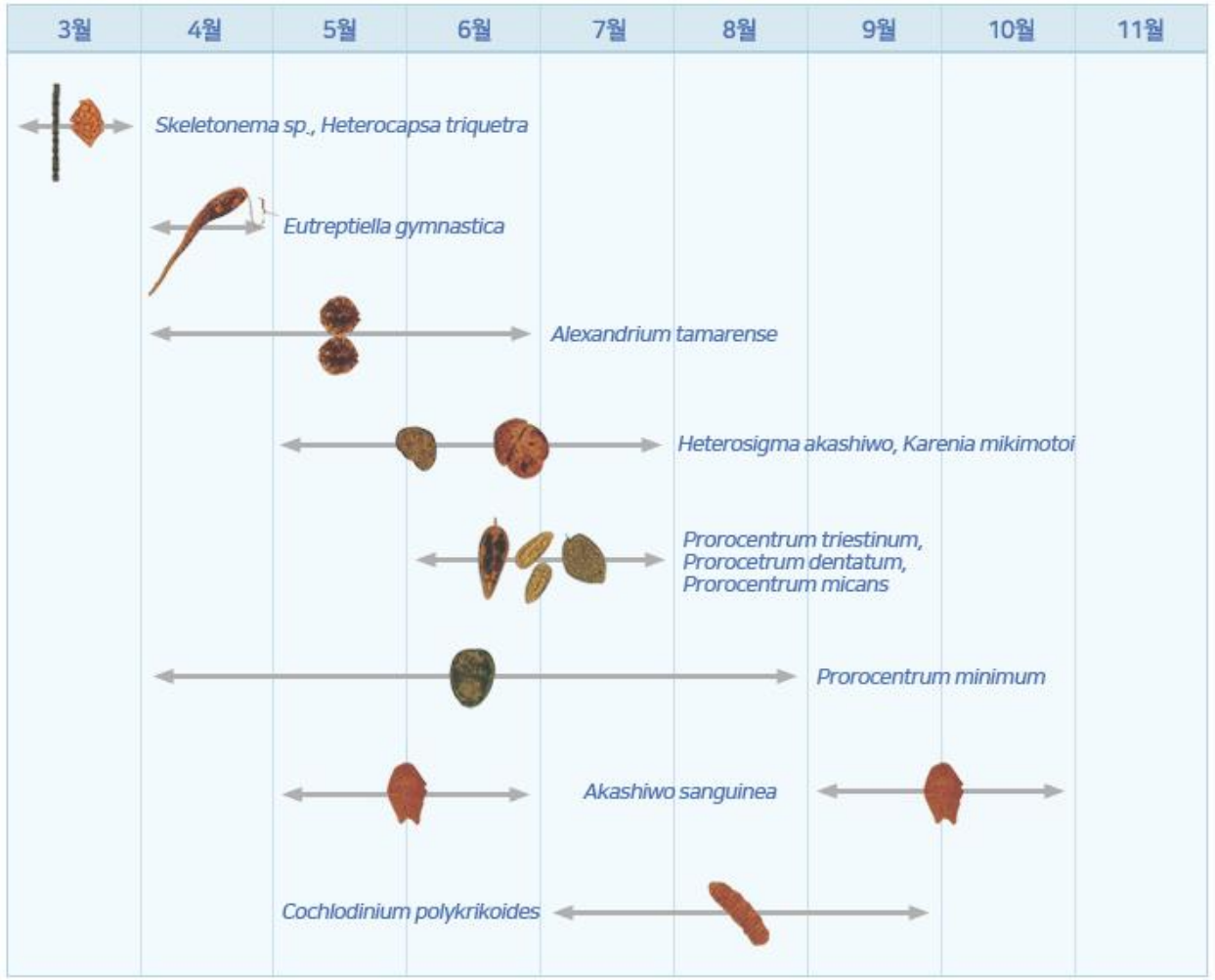




적조 현상이란, 식물 플랑크톤의 대량 번식으로 바닷물의 색깔이 적색, 황색, 적갈색으로 변색되는 자연 현상이다.
적조 현상은 꾸준히 수산업에 막대한 피해를 끼치고 있으며, 현재 해결책이 명확하지 않기 때문에 예방책을 마련하는 것이 중요하다. 따라서, 적조 현상이 발생하는 요소를 분석하고자 한다.

1 Background

Cause of Red Tide Phenomenon

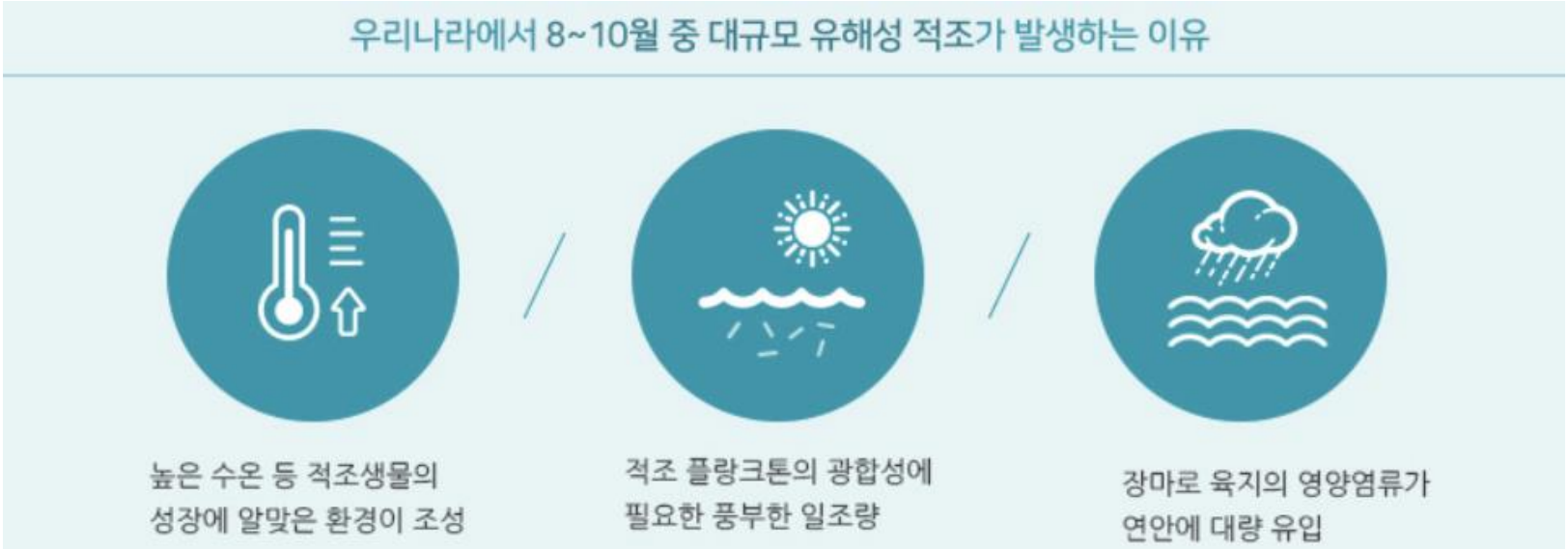


→ 적조 발생 식물성 플랑크톤

→ 적조 발생의 가장 큰 문제점: **"어류 폐사"**

1 Background

Cause of Red Tide Phenomenon



- 일반적으로 적조 현상 여름에 발생
  → 식물성 플랑크톤이 광학성하지 적절한 수온 및 일조량 제공
- 강수량이 증가하면 육지의 영양염류가 유입되기 때문에 부영양화 발생
  → 식물성 플랑크톤의 에너지원이 증가하기 때문에 개체수 증가

2 Dataset

About Dataset

기상/기후

적조발생 예측데이터

한국기상산업기술원

조회 140 | 다운로드 13

관심데이터

사용자 평점 5 평가하기

측정 기간: 2014년01월01일~2019년08월31일

가격	무료	배포주기	1년
생성날짜	2021-12-03	업데이트	2021-12-15
유형	ZIP	키워드	적조, 해양, 바다, 관측, 센서, 기상, 기후, 기상청, 예측, 화학
내용	본 데이터는 적조발생 예측데이터입니다. 확장자: csv 열람 활용방법: 범용프로그램을 통하여 데이터 열람 *sample 데이터 이므로 추가 필요시 담당자에게 연락 바람.		
활용예제	해양 적조발생 예측		
연락처	07050035024	이메일	center_bigdata@kmiti.or.kr
데이터 가공유형	원천		

출처: 환경 빅데이터 플렛폼 (https://www.bigdata-environment.kr/user/main.do)

2 Dataset

Variable Explanation

	변수	의미	설명
독립변수(X)	Temp	기온 (℃)	적조 발생 확률과 비례 관계
	WTemp	수온 (℃)	
	Wind	풍속 (m/s)	적조 발생 확률과 반비례 관계
	Rain	일강수량 (mm)	적조 발생확률과 비례 관계
	Salt	염도 (PSU)	최적의 적조 발생 환경: 32~33PSU
	PH	수소 이온 농도 지수	최적의 적조 발생 환경: PH 8
	DO	용존 산소량 (ppm)	적조 발생 확률과 비례 관계
	COD	화학적 산소 요구량 (ppm)	적조 발생 시, 증가
	Turbidity	탁도 (NTU)	적조 발생 시, 증가
	TN	총질소 (mg/L)	적조 발생 확률과 비례 관계
	TP	총인 (mg/L)	
종속변수(Y)	RedTide	적조 발생 여부: 0=발생X, 1=발생	-

3 Data Preparation

Data Verification

```
RedTide=read.csv('RedTide.csv')
str(RedTide)
# 'data.frame': 2068 obs. of  13 variables:
# $ Date     : num  20140101 20140102 20140103 20140104 20140105 ...
# $ WTemp    : num  7.7 7.7 7.7 7.8 7.7 7.6 7.6 7.5 7 6.8 ...
# $ Temp     : num  7.66 7.68 7.69 7.77 7.66 ...
# $ Rain     : num  0 0 0 0 0 0 0 4.2 0 0 ...
# $ Wind     : num  5.26 1.22 1.24 2.42 2.69 ...
# $ Salt     : num  29.3 29.3 29.3 29.2 29.3 ...
# $ DO       : num  10.07 10 9.98 9.95 9.89 ...
# $ COD      : num  1.168 1.297 1.163 0.845 0.818 ...
# $ PH       : num  7.84 7.86 7.85 7.85 7.87 ...
# $ Turbidity: num  262.7 194.5 86.7 68.6 85.9 ...
# $ TN       : num  0.29 0.29 0.29 0.29 0.29 ...
# $ TP       : num  0.0262 0.0262 0.0262 0.0262 0.0281 ...
# $ RedTide  : int  0 0 0 0 0 0 0 0 0 0 ...
```

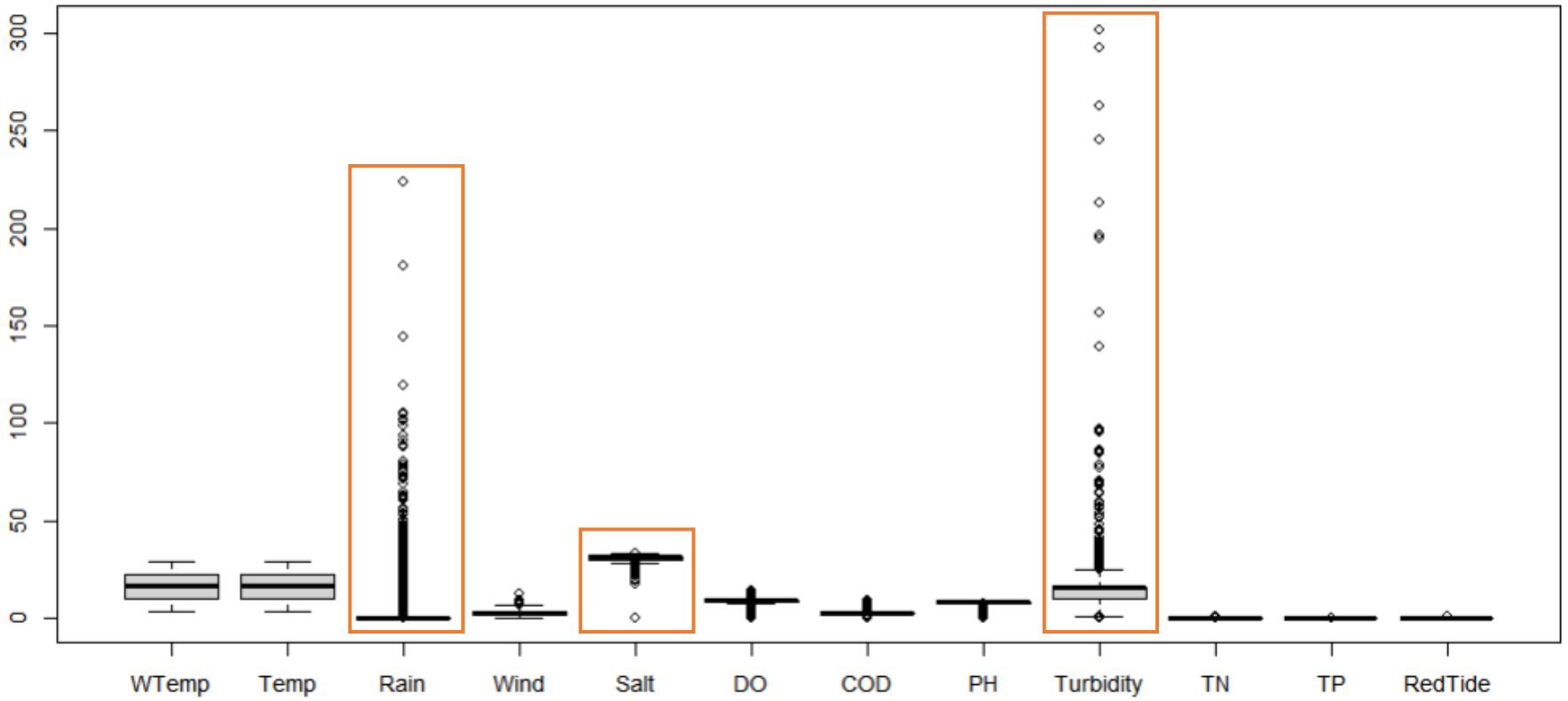
```
RT=subset(RedTide,subset=RedTide==1)
RT$Date # 보통 8~9월에 많이 발생
#  [1] 20140825 20140903 20140906 20140907 20140908 20140909
# [7] 20140910 20140911 20150810 20150811 20150812 20150813
# [13] 20150814 20150815 20150816 20150817 20150818 20150819
# [19] 20150820 20150821 20150822 20150831 20150901 20150902
# [25] 20150904 20150907 20150908 20150909 20150910 20150911
# [31] 20160820 20160821 20160822 20160823 20180802 20190827
# [37] 20190828
```

적조 현상 여름(8~9월)에만 발생 ←

3 Data Preparation

Missing Value Treatment

```
# 1) 결측치 확인 및 처리
sum(is.na(RedTide)) # 0 = 결측치 없음

# 2) 이상치 확인 및 제거
boxplot(RedTide[-1])
```



⇨ Rain, Salt, Turbidity
이상치 제거

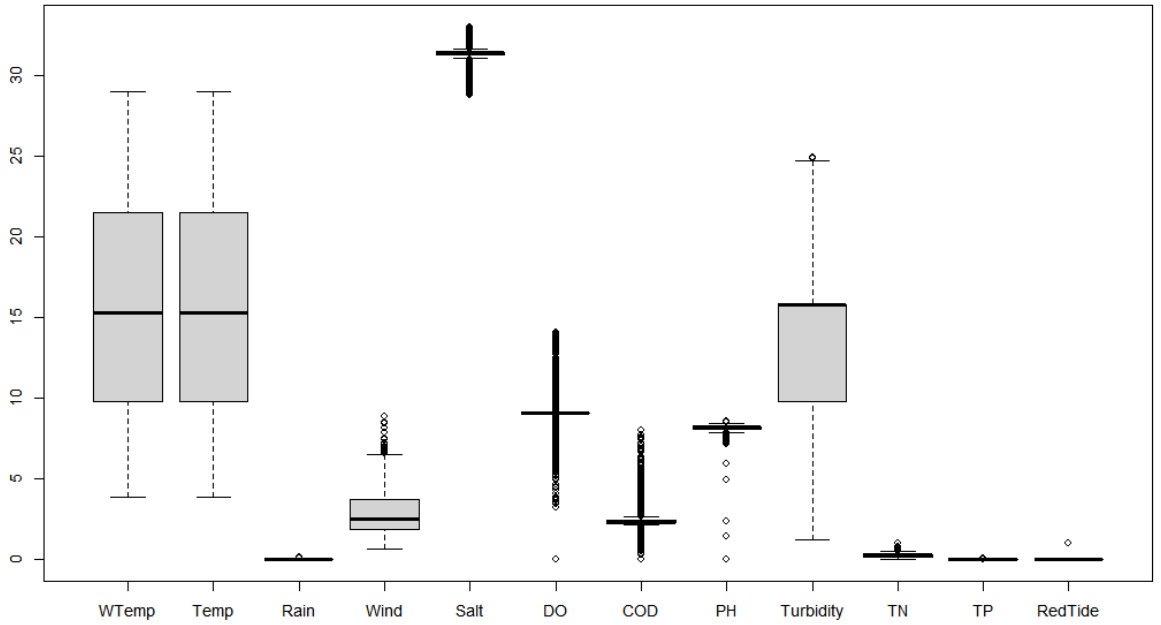
3 Data Preparation

Outlier Treatment

```
# Rain 이상치 제거
boxplot(RedTide$Rain)$stats # 0~0.2 → 하한값/상한값 확인
RedTide=subset(RedTide,0<=Rain & Rain<=0.2)
# Salt 이상치 제거
boxplot(RedTide$Salt)$stats # 28.78491~32.99567
RedTide=subset(RedTide,28.78491<=Salt & Salt<=32.99567)
# Turbidity 이상치 제거
boxplot(RedTide$Turbidity)$stats # 1~25.21942
RedTide=subset(RedTide,1<=Turbidity & Turbidity<=25) → 하한값/상한값 외 이상치 제거

# 차원확인
dim(RedTide) # [1] 1353   13 -> 약 700행 삭제
```

```
     [,1]
[1,]  2.1 → 상한값
[2,]  4.4
[3,]  5.4
[4,]  6.3
[5,]  7.9 → 하한값
```

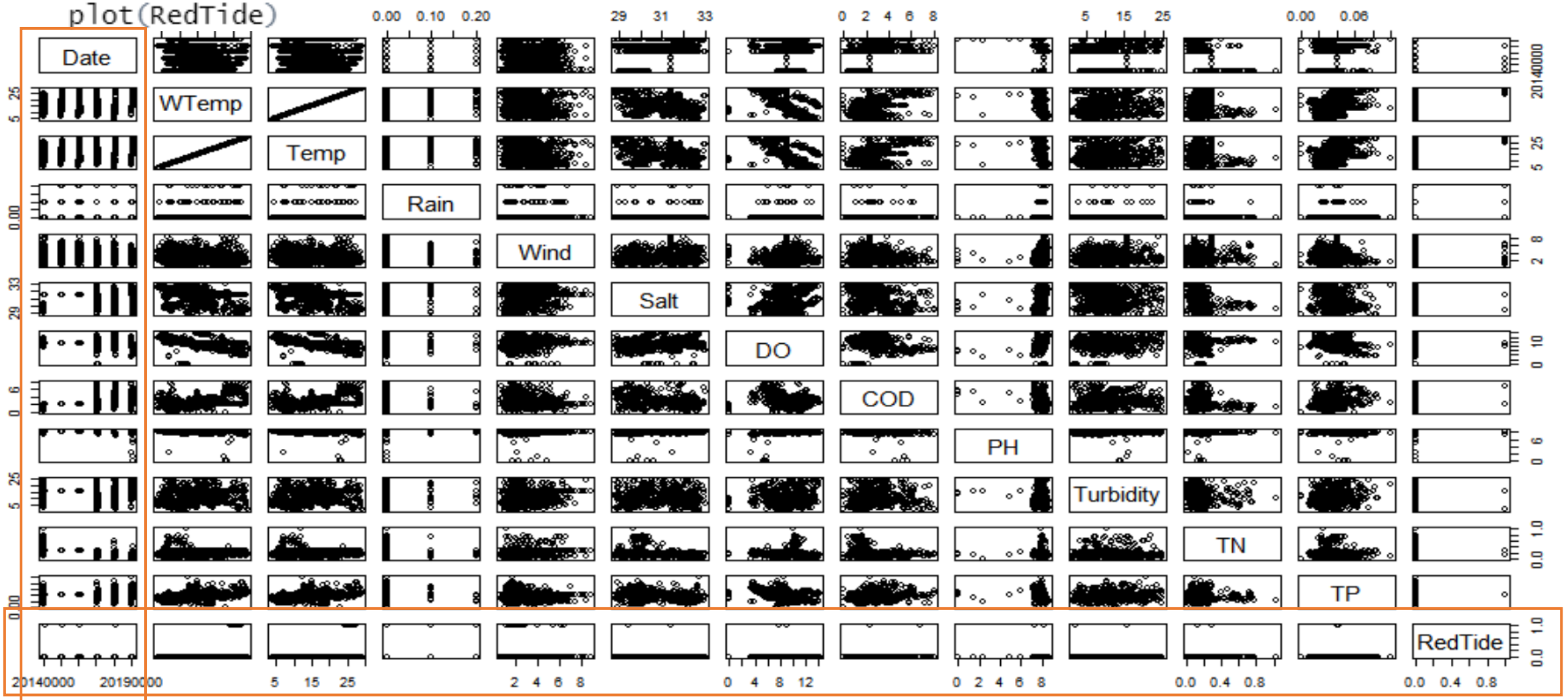
추가적으로 이상치를 제거하면 모든 행이
삭제되어 더 이상의 이상치 제거하지 않음



3 Data Preparation

Data Verification

```
# 데이터 확인
plot(RedTide)
```



4 Data Analysis

1) Correlation

```
RedTide2=RedTide[-1] # Date 제외
corr=cor(RedTide2) # 상관분석
#                WTemp        Temp        Rain        Wind        Salt          DO
# WTemp       1.00000000  0.99998944  0.019816550 -0.23461689 -0.23506338 -0.4052641633
# Temp        0.99998944  1.00000000  0.019797619 -0.23453756 -0.23515812 -0.4051372036
# Rain        0.01981655  0.01979762  1.000000000  0.07496405  0.01207697 -0.0082998948
# Wind       -0.23461689 -0.23453756  0.074964051  1.00000000  0.06747942  0.0598728959
# Salt       -0.23506338 -0.23515812  0.012076968  0.06747942  1.00000000  0.2445760612
# DO         -0.40526416 -0.40513720 -0.008299895  0.05987290  0.24457606  1.0000000000
# COD         0.33432530  0.33418235  0.012251692 -0.08711536 -0.28741758 -0.2742251117
# PH         -0.12537156 -0.12559070  0.026758737  0.04703072  0.28055175  0.2451105184
# Turbidity   0.17660030  0.17673580  0.058508837  0.04213868  0.07538034  0.0842605060
# TN          0.04687081  0.04691662  0.021441657  0.08551785 -0.16669651  0.0614530320
# TP          0.33342799  0.33345373  0.030603377 -0.08303982 -0.22322835 -0.2775321587
# RedTide     0.19973114  0.19958204  0.032469051 -0.02155956  0.02001962  0.0004668346
#                 COD          PH   Turbidity          TN          TP       RedTide
# WTemp       0.334325301 -0.12537156  0.17660030  0.04687081  0.33342799  0.1997311410
# Temp        0.334182353 -0.12559070  0.17673580  0.04691662  0.33345373  0.1995820441
# Rain        0.012251692  0.02675874  0.05850884  0.02144166  0.03060338  0.0324690508
# Wind       -0.087115362  0.04703072  0.04213868  0.08551785 -0.08303982 -0.0215595600
# Salt       -0.287417581  0.28055175  0.07538034 -0.16669651 -0.22322835  0.0200196235
# DO         -0.274225112  0.24511052  0.08426051  0.06145303 -0.27753216  0.0004668346
# COD         1.000000000 -0.21890561 -0.03795197 -0.28613573  0.18880582 -0.0081473581
# PH         -0.218905613  1.00000000  0.06787176  0.17850070 -0.05683173  0.0295731357
# Turbidity  -0.037951971  0.06787176  1.00000000  0.28960891  0.33874228  0.0652164617
# TN         -0.286135733  0.17850070  0.28960891  1.00000000  0.14390781  0.0654755314
# TP          0.188805816 -0.05683173  0.33874228  0.14390781  1.00000000  0.0372010196
# RedTide    -0.008147358  0.02957314  0.06521646  0.06547553  0.03720102  1.0000000000
# 상관관계가 0.9 이상인 것은 없음
```

독립변수(X)와 종속변수(Y) 사이 상관관계 없음

→ 로지스틱 회귀분석 X

→ "분류나무모델" O

4 Data Analysis

1) Correlation

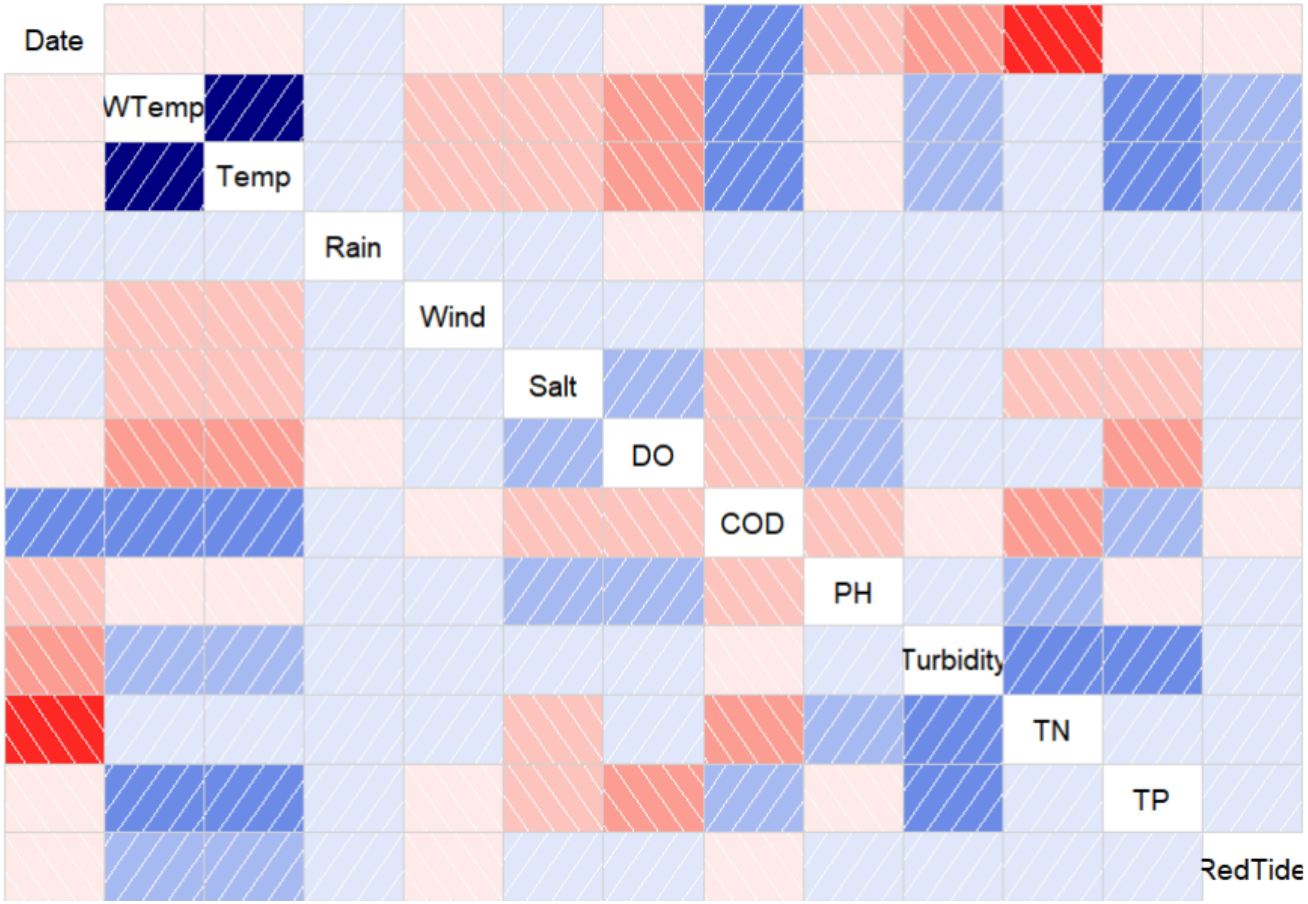
	변수	의미	상관분석 결과
독립변수(X)	Temp	기온 (℃)	0.9 이상: 매우 높은 상관성 있음
	WTemp	수온 (℃)	
	Wind	풍속 (m/s)	-
	Turbidity	탁도 (NTU)	-
	Salt	염도 (PSU)	-
	PH	수소 이온 농도 지수	-
	DO	용존 산소량 (ppm)	-0.27: 매우 낮은 상관성 있음
	COD	화학적 산소 요구량 (ppm)	
	Rain	일강수량 (mm)	0.02~0.14: 상관성 없음
	TN	총질소 (mg/L)	
	TP	총인 (mg/L)	
종속변수(Y)	RedTide	적조 발생 여부: 0=발생X, 1=발생	0.0004~0.19: 상관성 없음

→ Wtemp(수온)와 Temp(기온)를 제외하고 독립변수(X) 간 상관관계 없음

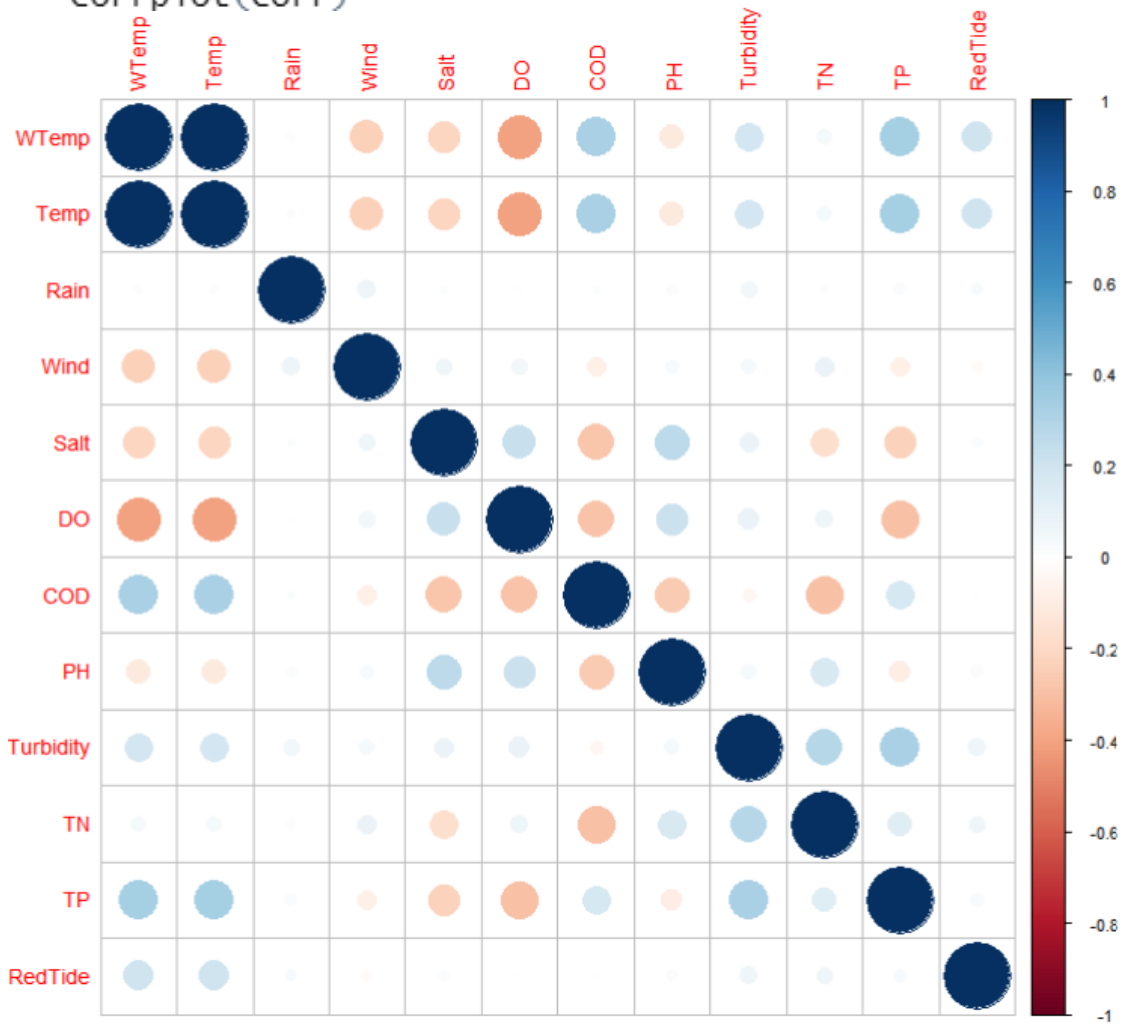
4 Data Analysis

1) Correlation

```
corrgram(RedTide)
```




```
corrplot(corr)
```

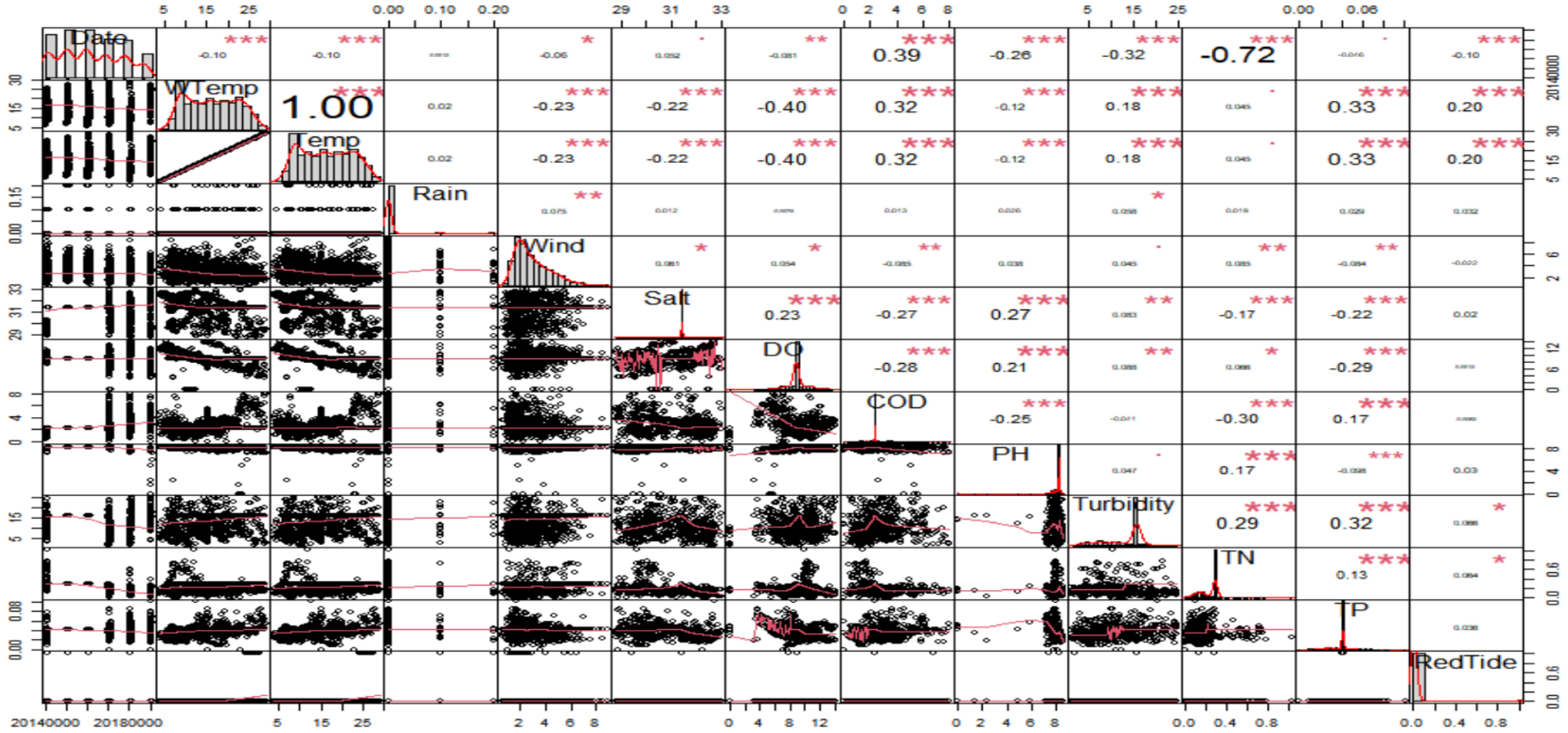


→ Wtemp(수온)와 Temp(기온)을 제외하고 강한 상관성 X

4 Data Analysis

1) Correlation

```
chart.Correlation(RedTide,histogram=T)
```



4 Data Analysis

2) Decision Tree

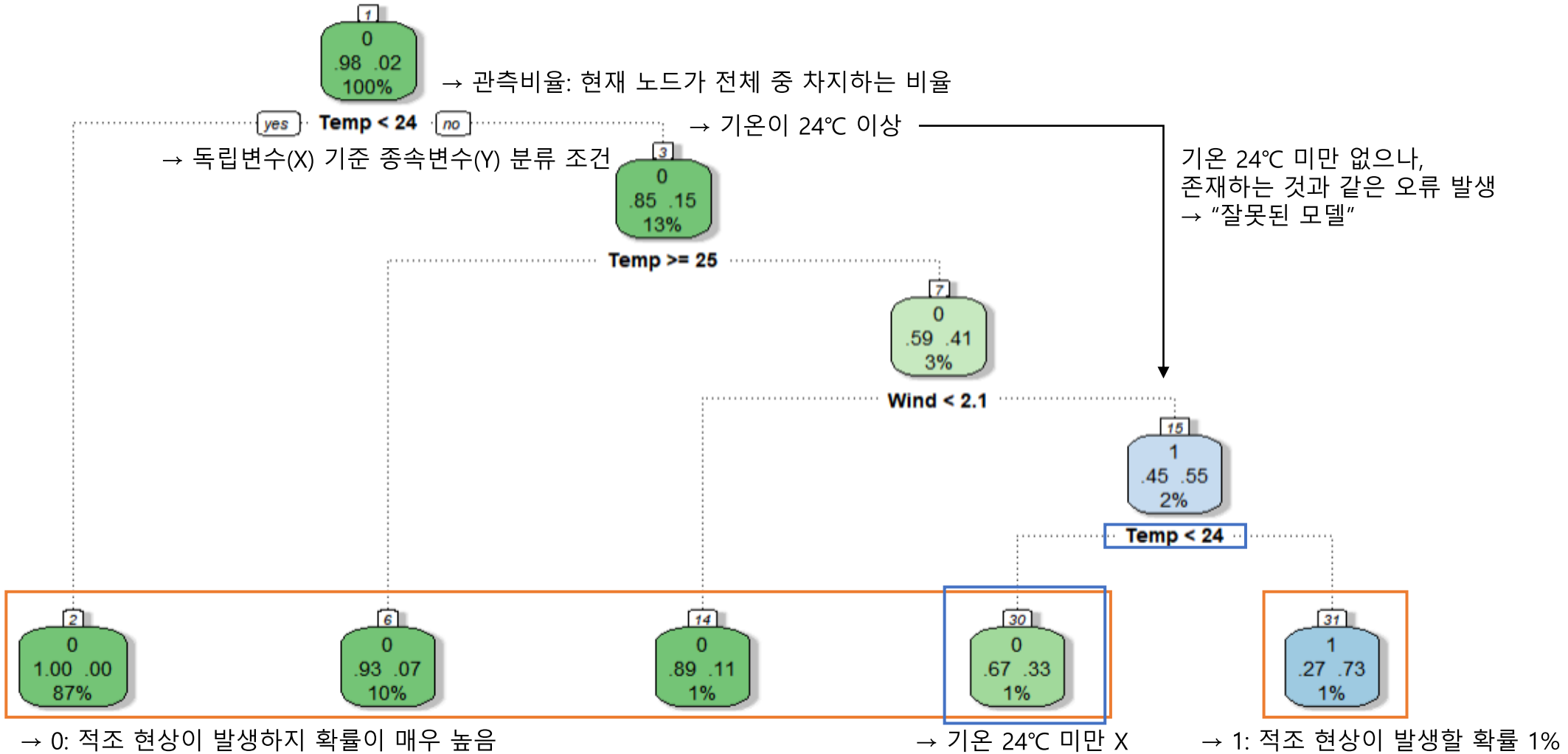
```
# 종속변수(Y) 범주 별 발생 빈도수 확인
# 0: 적조 발생하지 않음, 1: 적조 발생함
table(RedTide2$RedTide)
#    0    1
# 1326   27
prop.table(table(RedTide2$RedTide))
#          0          1
# 0.98004435 0.01995565   → 적조 발생하지 않을 확률: 0.98 매우 높음

RedTide2$RedTide=as.factor(RedTide2$RedTide) # 숫자형 -> 요인형 변환

# 샘플 추출
set.seed(123)
samp=sample(x=nrow(RedTide2),0.7*nrow(RedTide2)) → 훈련 데이터 70% 평가 데이터 30%
train=RedTide2[samp,]
test=RedTide2[-samp,]

# 분류나무모델 만들기
tree=rpart(formula=RedTide~.,data=train)
```

```
n= 947

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 947 19 0 (0.97993664 0.02006336)
   2) Temp< 24.06625 820  0 0 (1.00000000 0.00000000) *
   3) Temp>=24.06625 127 19 0 (0.85039370 0.14960630)
     6) Temp>=24.68229 98  7 0 (0.92857143 0.07142857) *
     7) Temp< 24.68229 29 12 0 (0.58620690 0.41379310)
      14) Wind< 2.124583 9  1 0 (0.88888889 0.11111111) *
      15) Wind>=2.124583 20  9 1 (0.45000000 0.55000000)
        30) Temp< 24.26125 9  3 0 (0.66666667 0.33333333) *
        31) Temp>=24.26125 11  3 1 (0.27272727 0.72727273) *
```

4 Data Analysis

2) Decision Tree

```
fancyRpartPlot(tree)
```



→ 관측비율: 현재 노드가 전체 중 차지하는 비율

→ 독립변수(X) 기준 종속변수(Y) 분류 조건

→ 기온이 24℃ 이상

기온 24℃ 미만 없으나,
존재하는 것과 같은 오류 발생
→ "잘못된 모델"

→ 0: 적조 현상이 발생하지 확률이 매우 높음

→ 기온 24℃ 미만 X

→ 1: 적조 현상이 발생할 확률 1%

4 Data Analysis

2) Decision Tree

```
# 모델 예측하기
test_pred=predict(tree,newdata=test,type='class')
test_real=test$RedTide
tab=table(test_real,test_pred)
tab
#           test_pred
# test_real   0   1
#         0 402   1
#         1   6   2

# 정밀도(Precision)=(TP)/(TP+FP)
preci=tab[2,2]/sum(tab[,2])
preci # 0.6666667
# 민감도(TPR)=(TP)/(TP+FN)
Recall=tab[2,2]/sum(tab[2,])
Recall # 0.75
# F-측정치(F-Measure)=(정밀도*민감도)/(정밀도+민감도)
F_Measure=(preci*Recall)/(preci+Recall)
F_Measure # 0.3529412
```

→ 분류모델에서 분류 예측한 범주와 실제 분류 범주를 교차 표 형태로 나타낸 혼동 생성

→ 종속변수(Y)가 하나의 범주로 몰려 있어 정확도가 아닌 F-측정치 계산

→ F-측정치가 0.35로 매우 낮기 때문에, 정확도가 매우 낮은 모델

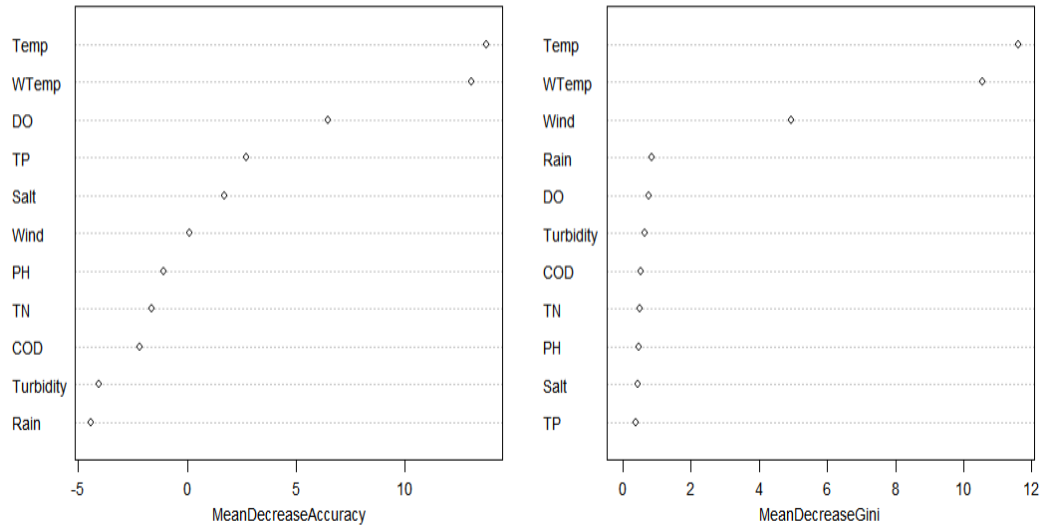
4 Data Analysis

3) Random Forest

```
sqrt(11) # 3.316625→ 모델의 독립변수(X) 개수 계산: 3개 혹은 4개

# 1) mtry=4
RF2=randomForest(formula=RedTide~.,data=RedTide2,ntree=500,mtry=4,
                 na.action=na.omit,importance=T)
RF2
# OOB estimate of  error rate: 2.22%→ 오분류율

# 2) mtry=3
RF=randomForest(formula=RedTide~.,data=RedTide2,ntree=500,mtry=3,
                 na.action=na.omit,importance=T)
RF
# OOB estimate of  error rate: 2.14%
```

→ 오분류율이 더 작은 독립변수(X) 개수 4개 선택

```
# 독립변수(X) 중요도 구하기
RF$importance
#                      0           1 MeanDecreaseAccuracy MeanDecreaseGini
# WTemp      1.608282e-02 0.017120618          1.583202e-02         6.8675472
# Temp       1.744924e-02 0.033676213          1.779576e-02         8.2253388
# Rain      -5.261013e-06 0.001688167          3.495514e-05         1.1145045
# Wind       1.830659e-03 0.004184443          1.816228e-03         4.5600324
# Salt       9.944316e-06 0.004648413          9.577612e-05         0.4230589
# DO         1.579235e-03 0.006572944          1.661834e-03         0.5165968
# COD       -2.053870e-04 0.005280880         -1.155815e-04         0.4163729
# PH        -2.360622e-04 0.003176984         -1.629265e-04         0.3441054
# Turbidity -2.754429e-04 0.002644444         -2.123009e-04         0.7156425
# TN        -5.177153e-04 0.014137590         -2.382106e-04         0.3238264
# TP         1.092028e-05 0.004876912          1.081763e-04         0.2736021
varImpPlot(RF)
```



→ 중요 독립변수(X): Temp > WTemp > Wind > DO …

5 Conclusion

As a Result…

- 해당 분석을 통해 생성된 모델은 "**잘못된 모델**"
- 잘못된 모델인 이유:

① 이론적으로 종속변수(Y)와 상관성 있는 독립변수(X), 상관성이 없는 것으로 분석

② 종속변수(Y)를 결정하는 중요 일조량 등 독립변수(X) 없음

③ '1: 적조 현상 발생'를 결정할 종속변수(Y) 부족 (측정 기간 5년 이하로 짧음)

→ '0: 적조 현상 발생하지 않음' 출현 비율 98% 이상

- 따라서, **적조 현상이 발생하는데 기여하는 독립변수(X)**

및 적조 현상이 발생하는 데이터 추가 필요

THANK YOU