동국대학교
데이터마이닝
2022-1학기
기말 프로젝트

전복 성장 데이터를 활용한 양식업 효율 향상 연구

팀장: 김기호 2018110475
팀원: 김현 2017112282
배정민 2018110476
양승엽 2018110499

목차

1. 분석배경
2. EDA
3. Modeling
4. 결과해석

1. 분석 배경 : 전복의 성장 자료를 이용한 예측 및 양식 환경 효율 극대화

전복 양식 효율성 향상에 대한 꾸준한 관심

13년 연구 끝에...마침내 전복 양식기간 6개월 단축

방준호 기자 +구독

3년 걸린 양식 2년6개월로
품종개발·검증 10년 넘게 반복
해수부 "전복 값 낮아질 것"

전복의 성장 데이터로
판매 관리 & 양식 환경 관리를 한다면 어떨까?

2. EDA 데이터 소개

abalone_data

OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
2	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
3	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
4	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
...	...	...	...	...	...	...	...	...	...
4174	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10
4175	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9
4176	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10
4177	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12

4177 rows × 9 columns

2. EDA 데이터 소개

변수명	타입	설명	단위
Sex	nominal	M:수컷 F:암컷 I:치패(성별X)	
Length	continuous	껍질 중 가장 길이가 긴 부분의 길이	mm
Diameter	continuous	Length 측정 선 기준으로 수직길이	mm
Height	continuous	전복의 두께	mm
Whole weight	continuous	전복의 전체 무게	g
Shucked weight	continuous	전복 살의 무게	g
Viscera weight	continuous	건조된 전복의 내장 무게	g
Shell weight	continuous	전복 껍질의 무게	g
Rings	integer	전복 운문(나이테) 개수	

original data source: Marine Resources Division
data received: 1995

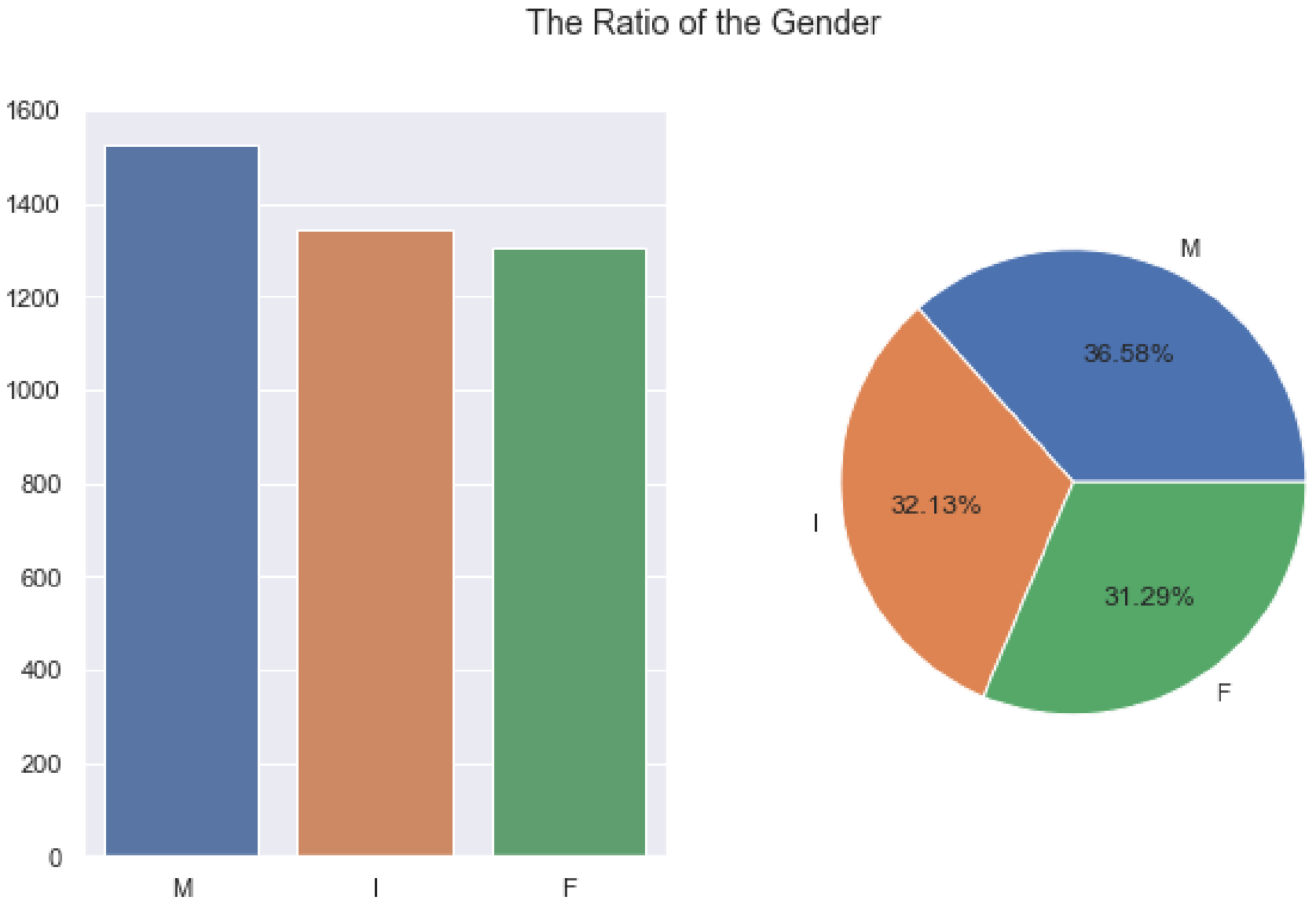
2. EDA 변수 단위 변환

연속형 자료 × 200

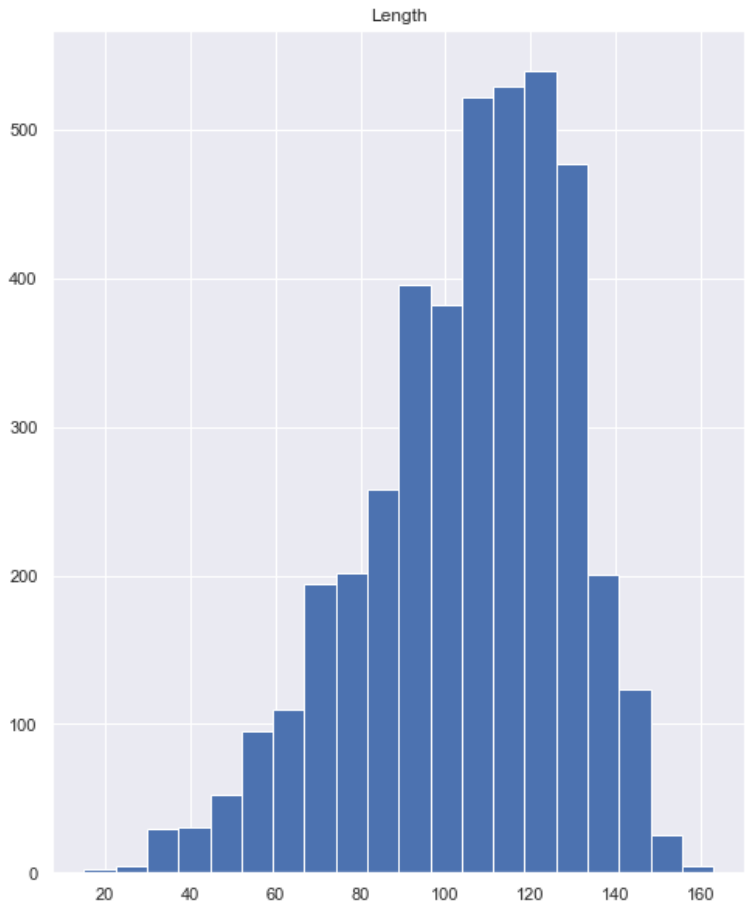
OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1	M	91	73	19	102.8	44.9	20.2	30	15
2	M	70	53	18	45.1	19.9	9.7	14	7
3	F	106	84	27	135.4	51.3	28.3	42	9
4	M	88	73	25	103.2	43.1	22.8	31	10
...	...	...	...	...	...	...	...	...	...
4174	M	118	88	27	193.2	87.8	42.9	52.1	10
4175	M	120	95	41	235.2	105.1	57.5	61.6	9
4176	F	125	97	30	218.9	106.2	52.5	59.2	10
4177	M	142	111	39	389.7	189.1	75.3	99.0	12

4177 rows × 9 columns

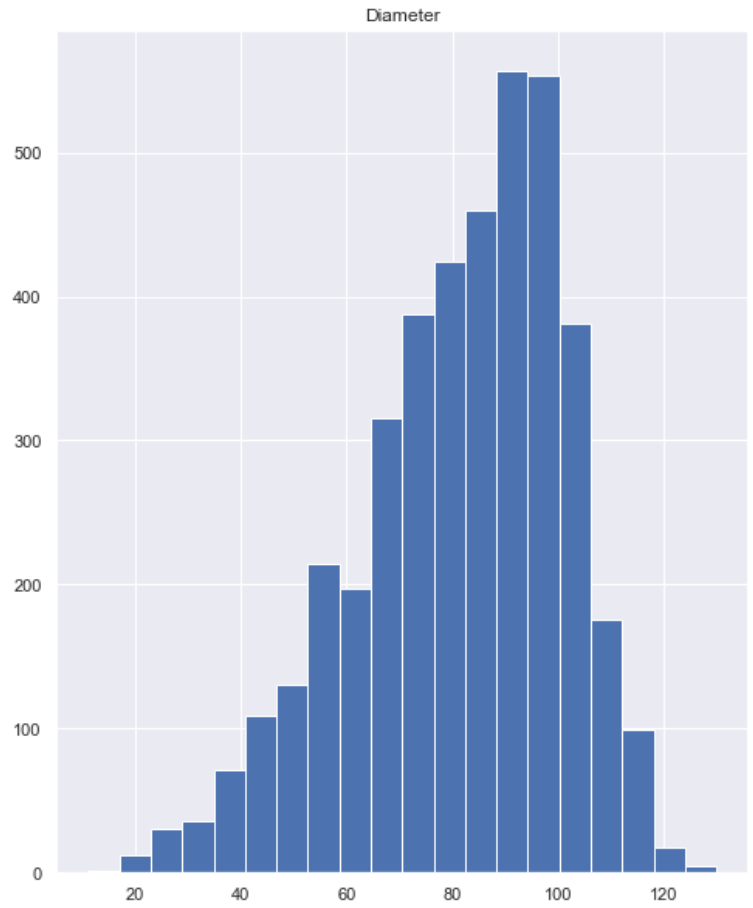
2. EDA 변수 별 분포 확인



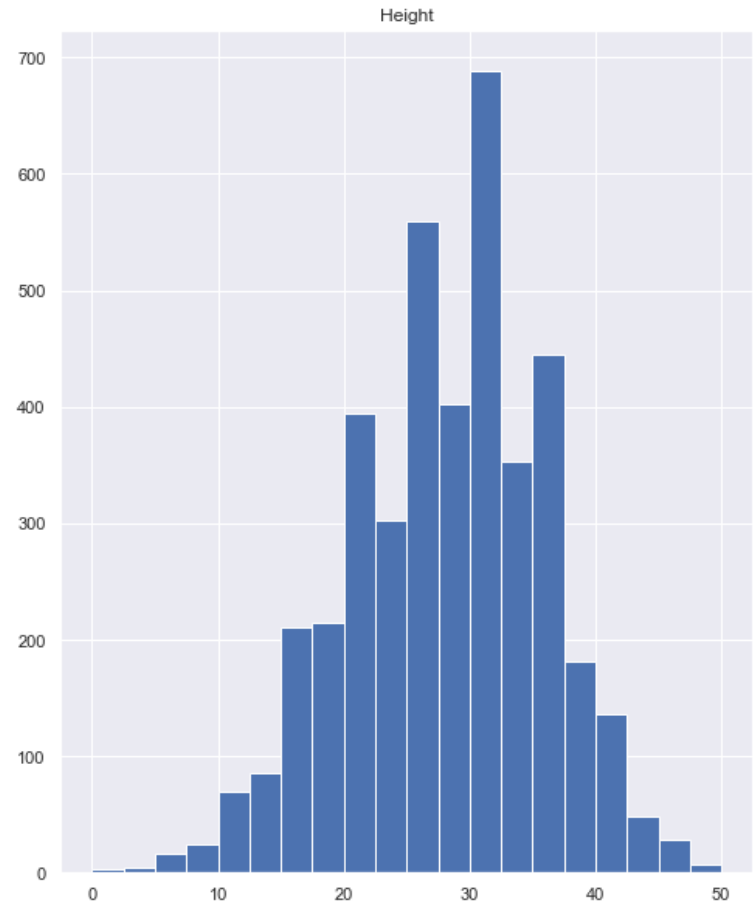
2. EDA 변수 별 분포 확인



Length



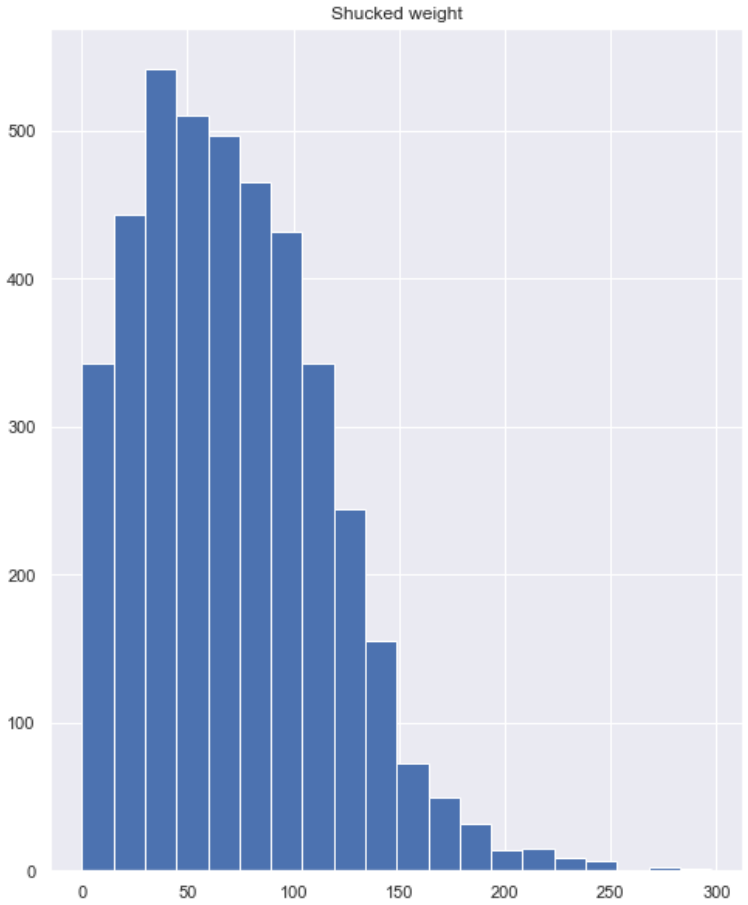
Diameter



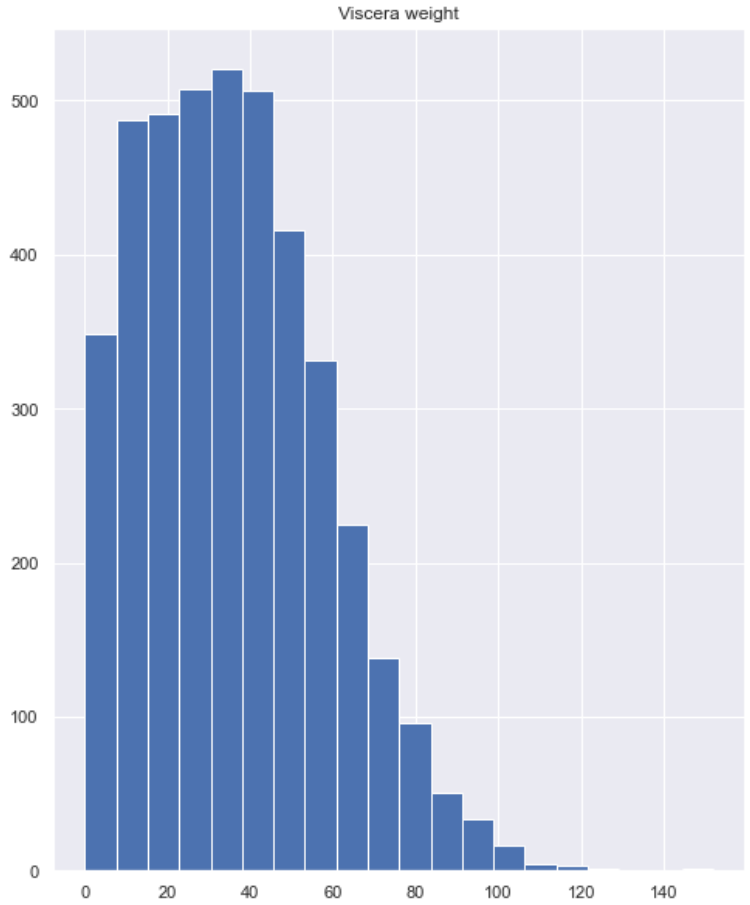
Height

길이 관련 변수들의 분포

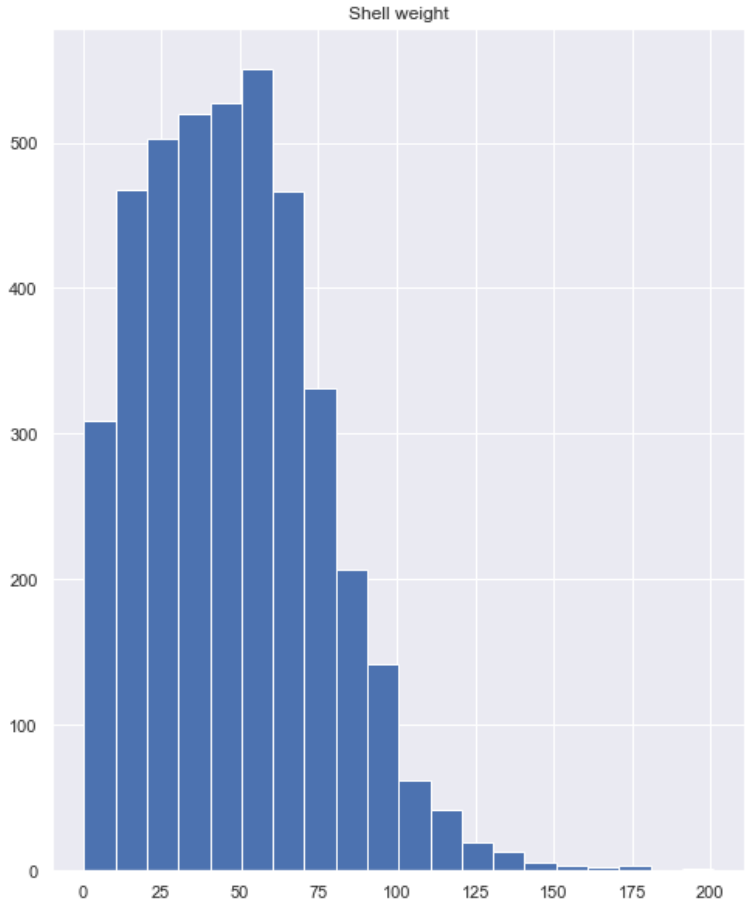
2. EDA 변수 별 분포 확인



Shucked weight



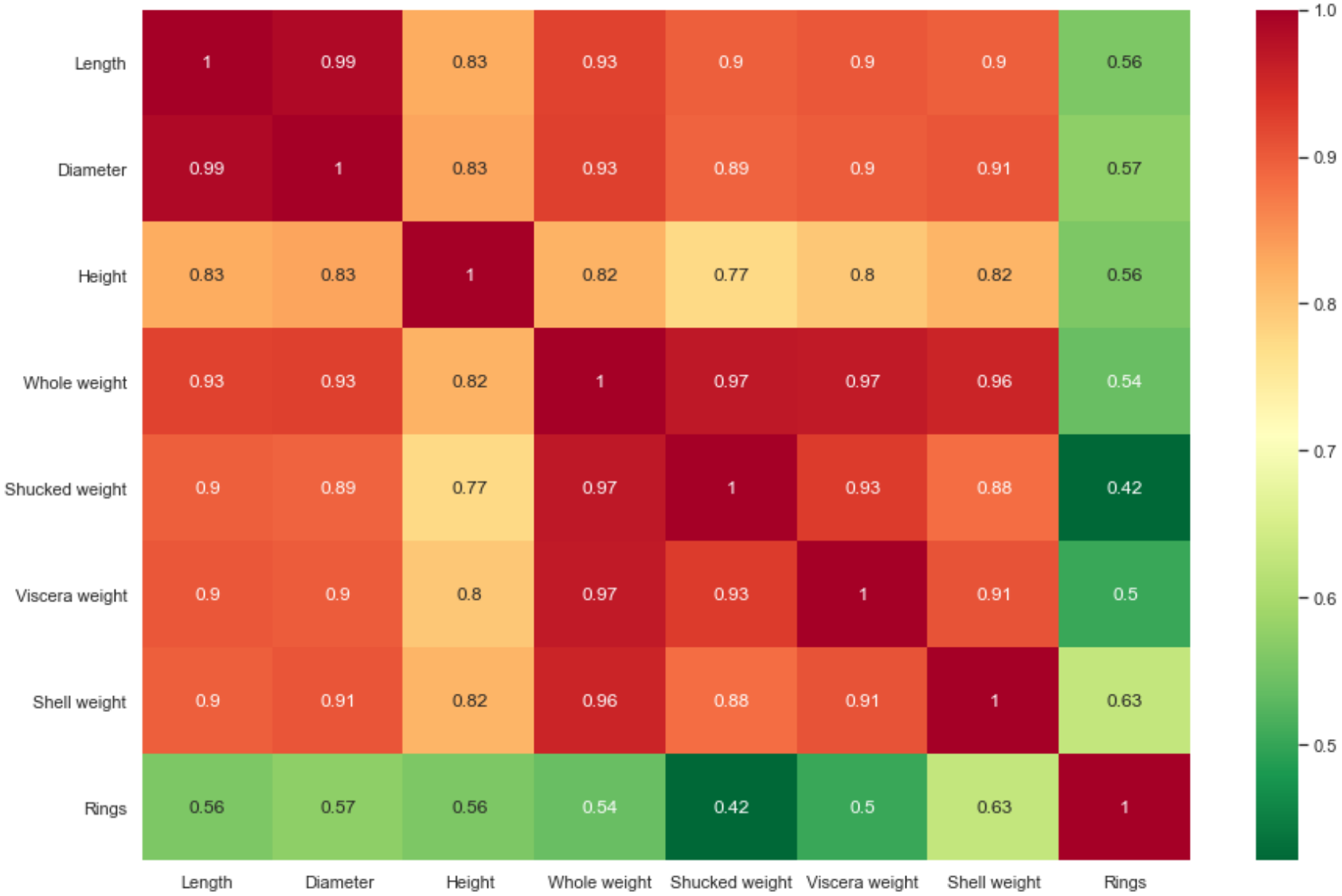
Viscera weight



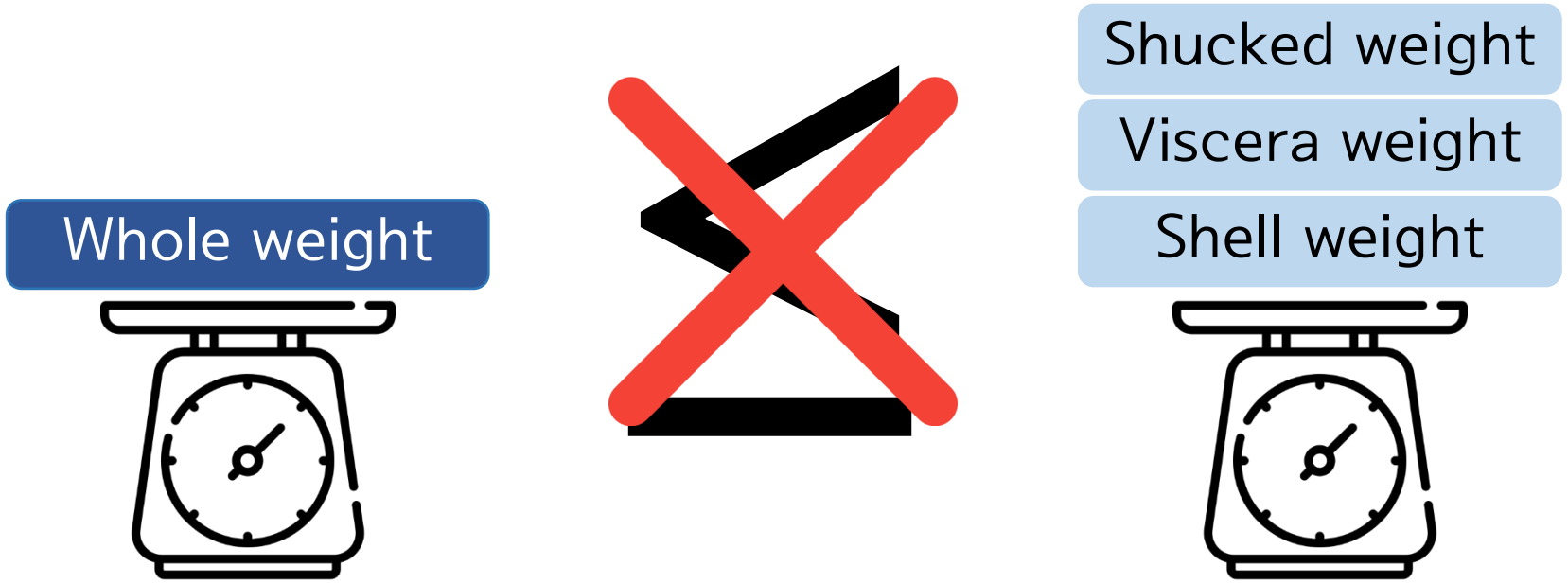
Shell weight

무게 관련 변수들의 분포

2. EDA 상관계수 확인


Length 1 0.99 0.83 0.93 0.9 0.9 0.9 0.56
Diameter 0.99 1 0.83 0.93 0.89 0.9 0.91 0.57
Height 0.83 0.83 1 0.82 0.77 0.8 0.82 0.56
Whole weight 0.93 0.93 0.82 1 0.97 0.97 0.96 0.54
Shucked weight 0.9 0.89 0.77 0.97 1 0.93 0.88 0.42
Viscera weight 0.9 0.9 0.8 0.97 0.93 1 0.91 0.5
Shell weight 0.9 0.91 0.82 0.96 0.88 0.91 1 0.63
Rings 0.56 0.57 0.56 0.54 0.42 0.5 0.63 1
Length Diameter Height Whole weight Shucked weight Viscera weight Shell weight Rings
1.0 0.9 0.8 0.7 0.6 0.5

2. EDA 이상 관측치 제거 : 무게



만약 (Shucked + Viscera+ Shell) Weight 가 Whole weight 보다 크거나 같다면
논리적 오류 발생

2. EDA 이상 관측치 제거 : 무게

OBS	Sex	…	Whole weight	Shucked weight	Viscera weight	Shell weight	…	(Shucked + Viscera + Shell) weight
43	M	…	14.0	6.3	4.7	4	…	15.0
44	M	…	8.4	5.1	3.0	2.4	…	10.5
45	F	…	8.4	3.5	2.5	3.0	…	9.0
46	M	…	40.6	17.5	9.0	15.0	…	41.5
…	…	…	…	…	…	…	…	…
3970	M	…	55.4	33.1	12.5	16.4	…	62
3997	M	…	26.8	11.5	5.7	70.1	…	87.3
4047	F	…	133.1	57.0	29.8	53.8	…	140.6
4144	M	…	271.8	128.4	65.1	81.0	…	274.5

→ 총 161개 관측치 제거

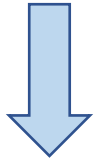
2. EDA 이상 관측치 제거 : 두께

이상치 대체

OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1258	I	86	68	0	85.3	41.3	17.2	23.0	8

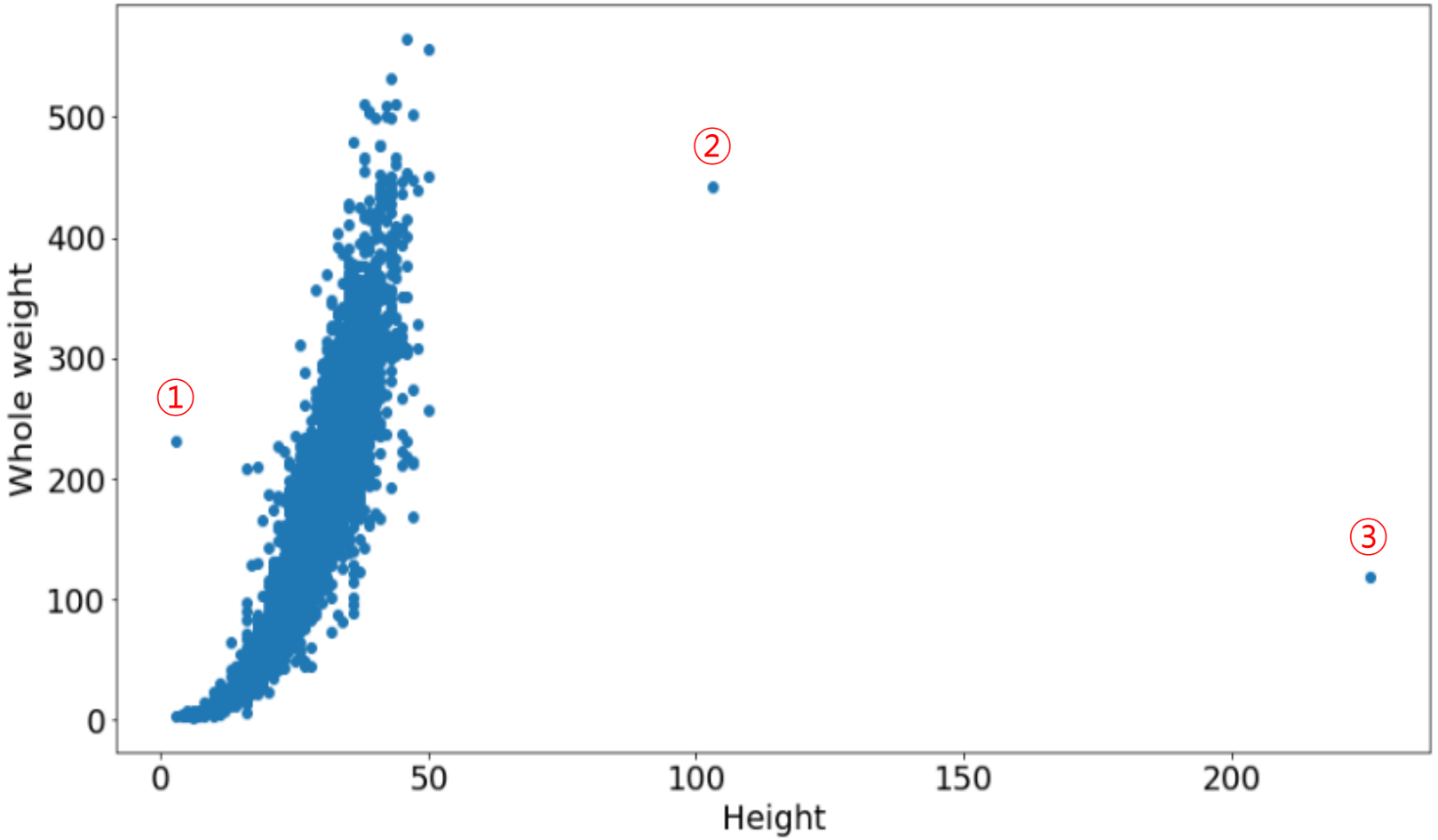
Height 변수와 가장 상관계수가 높았던 Diameter 변수를 이용

같은 Diameter 값을 갖는 모든 관측치들의 Height 평균으로 대체



OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1258	I	86	68	21.6	85.3	41.3	17.2	23.0	8

2. EDA 이상 관측치 제거 : 두께



분포 확인 과정에서 이상 의심 관측치 3개 발견

2. EDA 이상 관측치 제거 : 두께

①

OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1175	**F**	**127**	**99**	**3**	**231.3**	**102.3**	**61.6**	**57.7**	**9**
2170	I	33	23	3	2.9	1.1	0.6	1	2

두께(Height)가 3 이하인 관측치들

25	F	123	96	33	232.3	102.6	60.27	61	10
105	M	121	94	32	234.7	99.5	48.1	69	12
…	…	…	…	…	…	…	…	…	…
1175	**F**	**127**	**99**	**3**	**231.3**	**102.3**	**61.6**	**57.7**	**9**
…	…	…	…	…	…	…	…	…	…
4056	F	129	100	30	93.5	93.5	67.1	62	9

비슷한 전체 무게(Whole weight)를 기준으로 살핀 두께(Height) → 총 1개 관측치(1175) 제거

2. EDA 이상 관측치 제거 : 두께

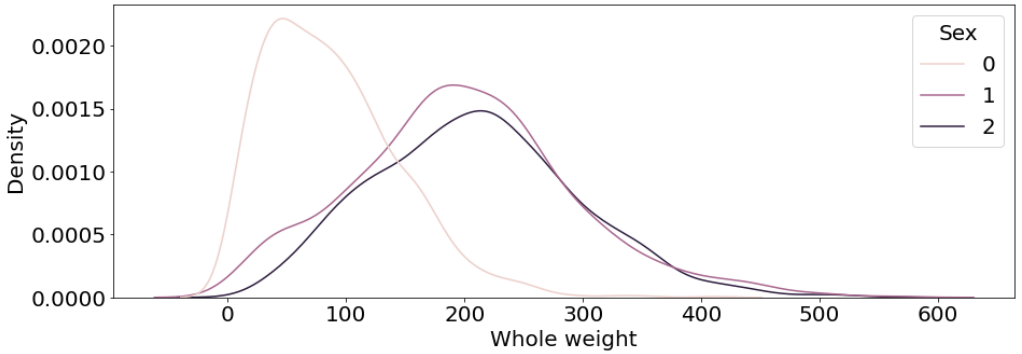
②+③

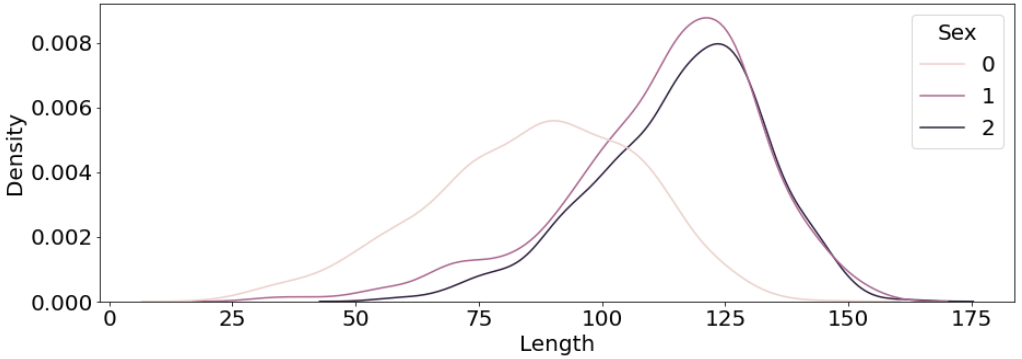
OBS	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
1	M	91	73	19	102.8	44.9	20.2	30	15
…	…	…	…	…	…	…	…	…	…
1418	M	141	113	103	442	221.5	97.3	102.4	10
…	…	…	…	…	…	…	…	…	…
2052	F	91	71	226	118.8	66.4	23.2	26.7	8
…	…	…	…	…	…	…	…	…	…
4177	M	142	111	39	389.7	189.1	75.3	99.0	12

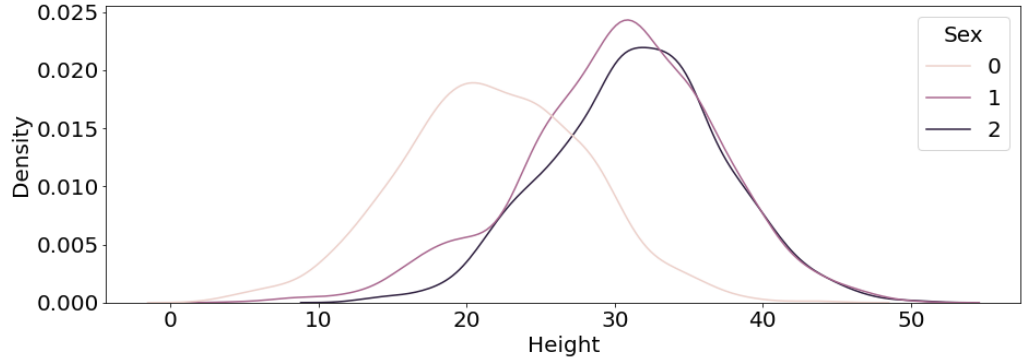
전체 무게(Whole weight)에 비해 두께(Height)가 극단적으로 큼

→ 총 2개 관측치(1418, 2052) 제거

2. EDA 성별 재범주화







0 : I
1 : M
2 : F

성별의 범주 별 특징을 파악

→ I와 M&F의 양상이 다름

2. EDA 성별 재범주화

성별에 따라 새로운 변수 'Adult' 생성 후 재범주화

Sex : I → 0
Sex : M&F → 1

OBS	Sex
1	M
2	M
3	F
4	M
5	I
…	…
4012	M

OBS	Adult
1	1
2	1
3	1
4	1
5	0
…	…
4012	1

2. EDA 파생변수 생성

다중공선성 확인

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight
VIF	41.690171	42.754963	6.774577	143.285733	35.667654	19.420346	26.192182

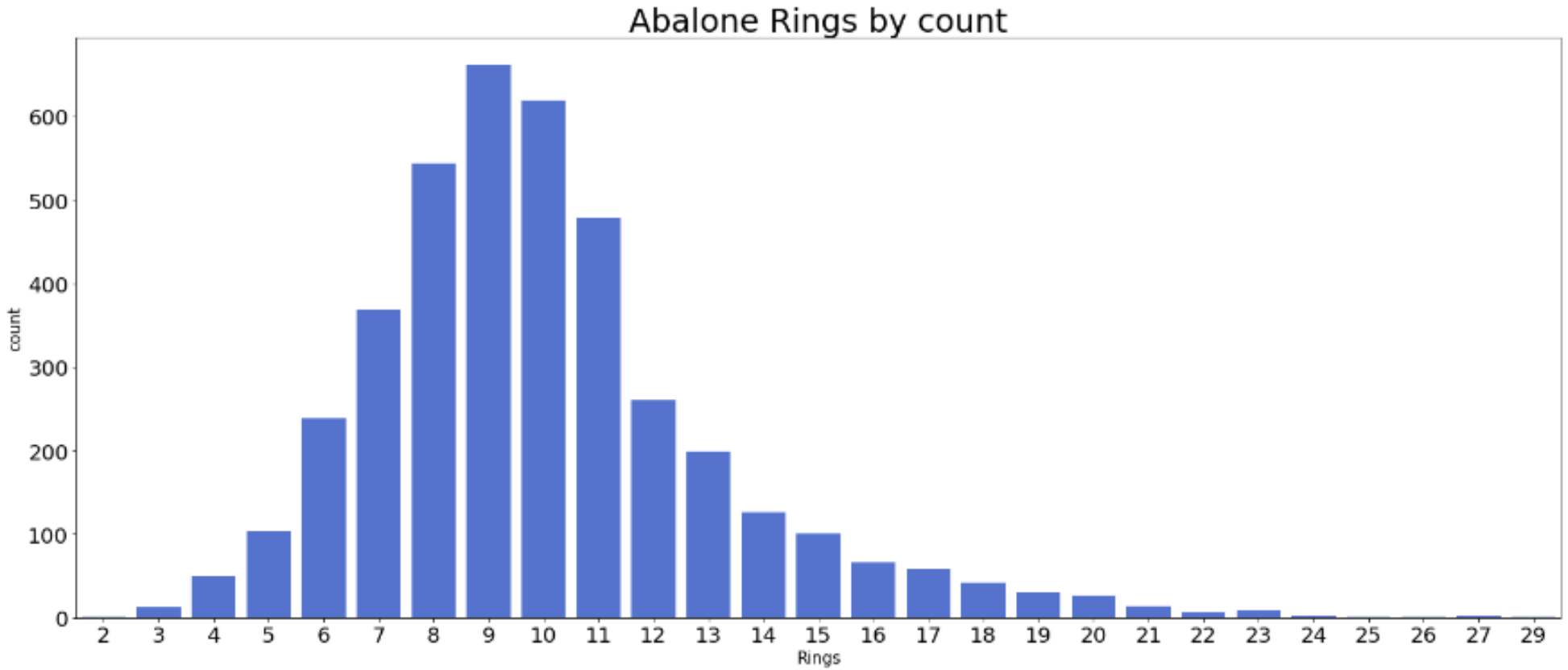
'Height' 이외의 변수에서 VIF값이 10보다 크게 나옴

대부분의 변수가 무게, 길이, 크기 관련 변수이므로 다중공선성 문제가 발생할 수 밖에 없음

파생변수 목록

변수명	설명	계산식
Volume	전복의 부피(반구 부피 기준)	$\frac{2}{3} \times \pi \times \frac{Length}{2} \times \frac{Diameter}{2} \times \frac{Height}{2}$
Moisture	전복의 수분 포함량	Whole weight - (Shucked + Viscera + Shell) weight
Area	전복의 수평 단면적	$\frac{Length}{2} \times \frac{Diameter}{2} \times \pi$
Density	전복의 단위 부피당 무게(밀도)	$\frac{Shucked\ weight}{Volume}$

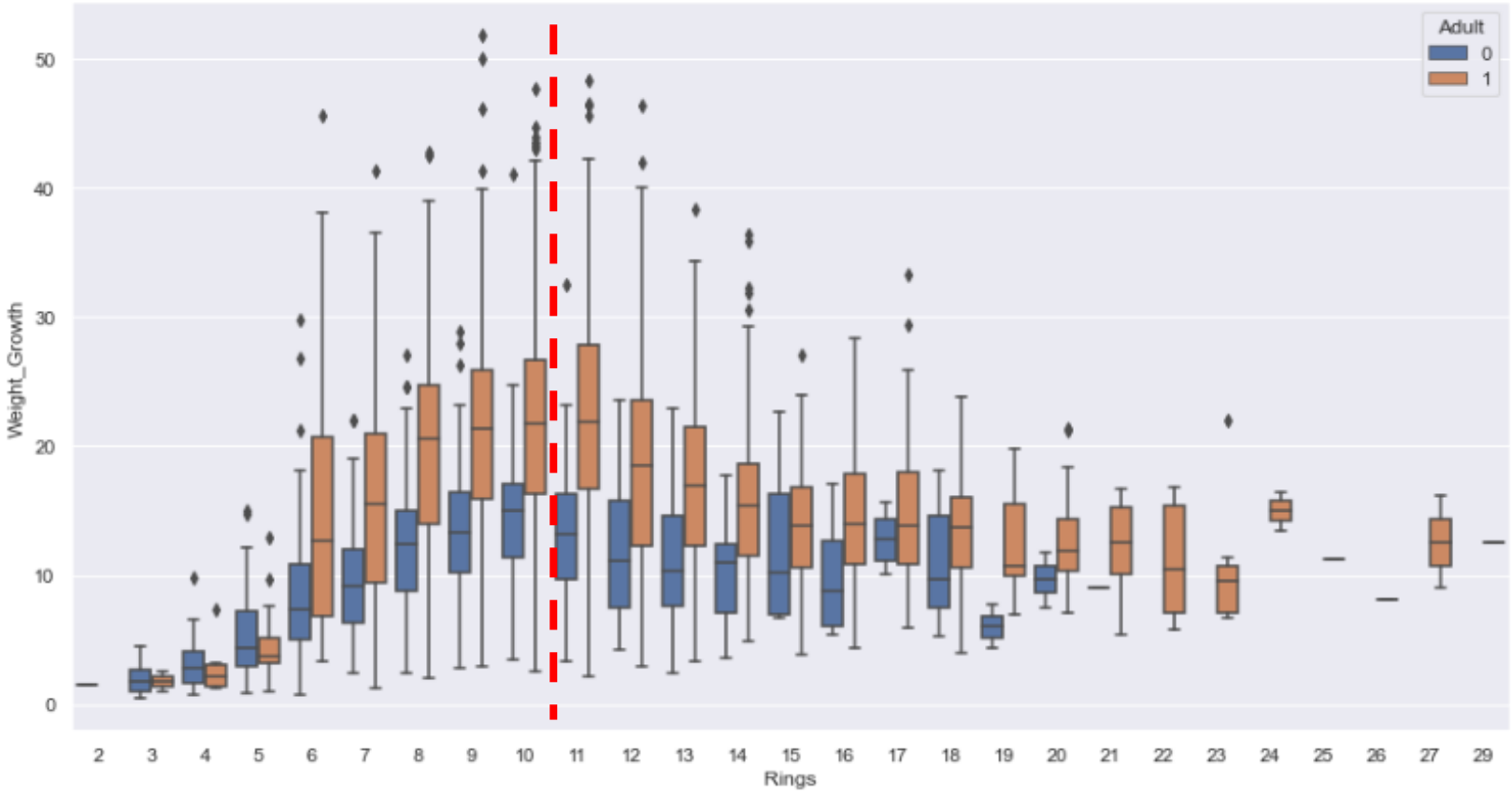
2. EDA 파생변수 생성



'Rings' 변수의 특징을 파악하고 분석하기 어려움

→ 무게 평균 성장률($\frac{Whole\ weight}{Rings}$)을 통해 재범주화

2. EDA 파생변수 생성

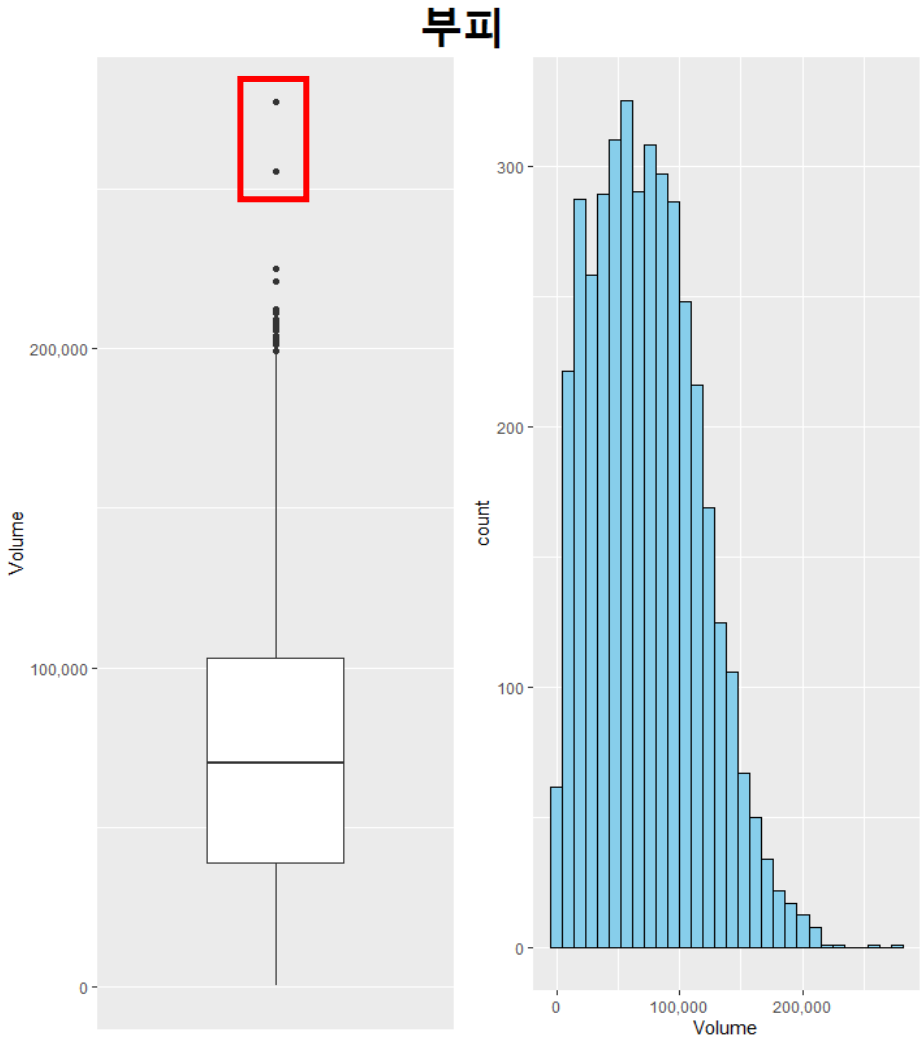


‘Rings’가 10까지는 무게 평균 성장률이 상승세, 그 이후로는 하락세

성장률의 추세를 통해 새로운 파생변수 ‘Status’ 생성

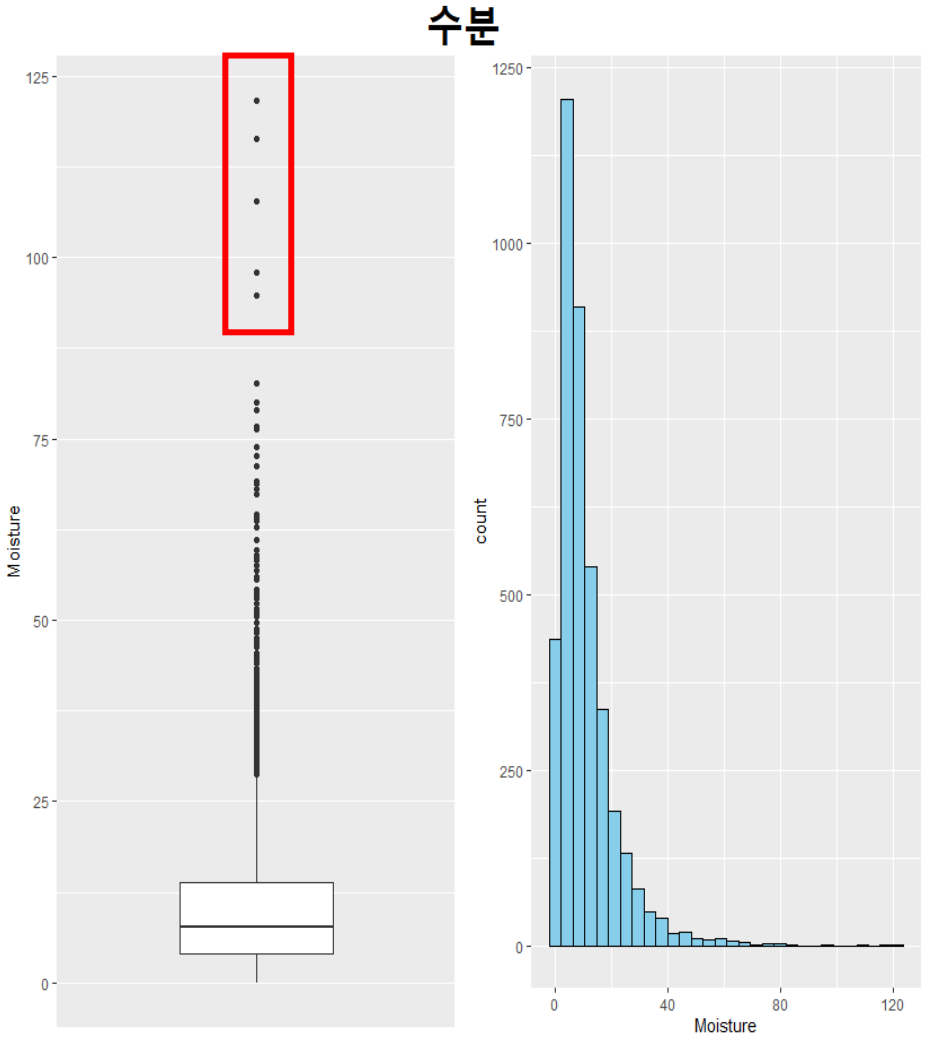
Rings : 0 ~ 10 → 0
Rings : 11 ~ 29 → 1
※ Adult 0 기준

2. EDA 파생변수 확인



큰 쪽에서 이상치 형성
→ 납득할 정도의 크기와 무게를 가진 개체들이라 제거X

2. EDA 파생변수 확인

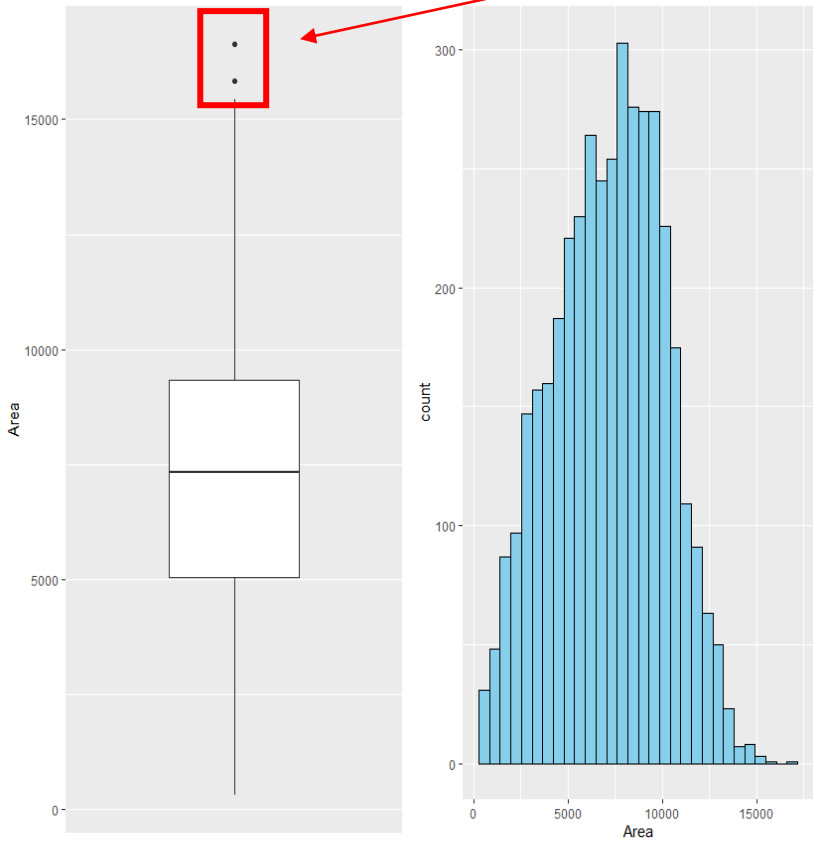


큰 쪽에서 이상치 형성
→ 자료의 불균형이 심하고 다른 변수 값이 클 때,
수분량도 많은 경우가 존재해 제거X

2. EDA 파생변수 확인

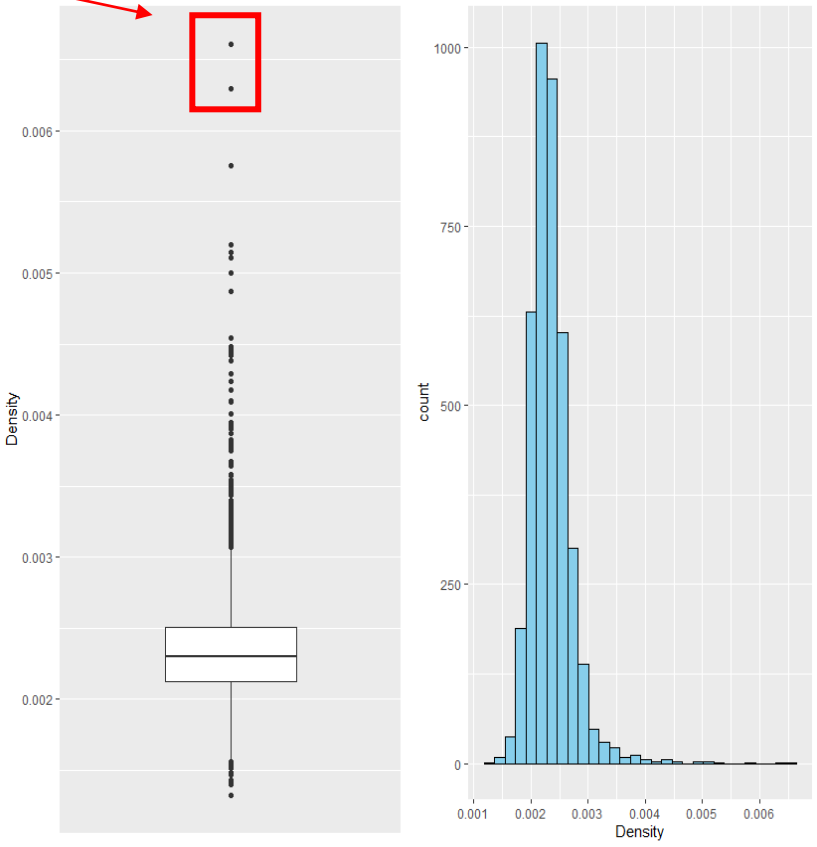
같은 개체

전복의 수평면적



큰 쪽에서 이상치 형성
→ 다른 전복들보다 비정상적으로 큼
총 2개 관측치 제거

밀도



2. EDA 변수 선택

파생변수 다중공선성 확인

	Adult	Volume	Moisture	Density
VIF	4.559843	7.026980	3.305360	3.632451

모든 파생변수에서 VIF값이 10보다 작게 나옴

- 파생된 변수에 원 종속 변수들 모두 사용
- 다중 공선성을 최소화 함으로서 모델 불안정성 최소화

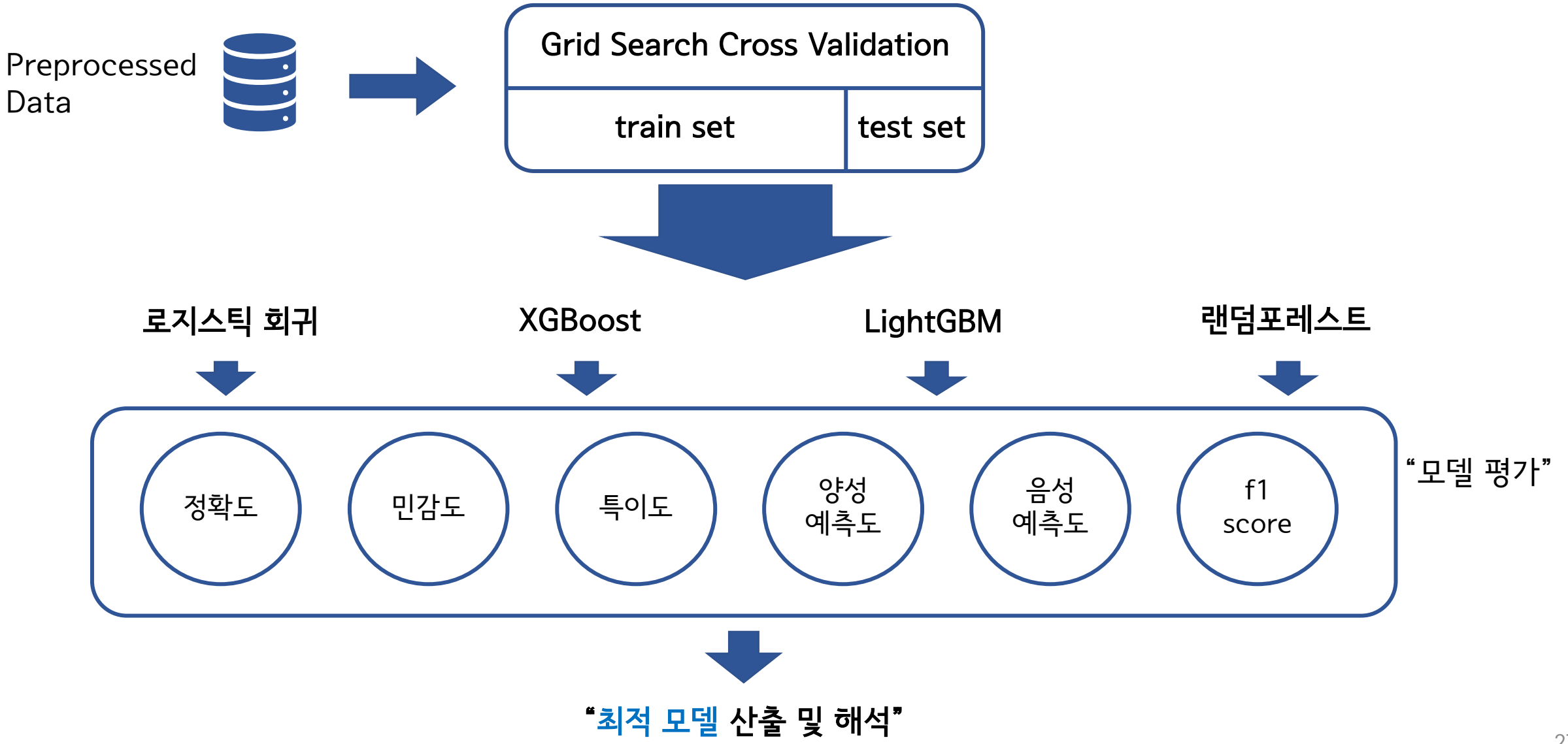
→ 위의 4개 변수로 분류 모델 적합 결정

2. EDA 데이터 변환

OBS	Adult	Volume	Moisture	Density
1	1	-0.934644	-0.283125	2.088153
2	0	-1.293239	-0.861550	0.639543
3	0	-0.245708	0.285971	-0.528174
4	0	-0.727205	-0.413737	0.298404
...	...	...	...	...
4007	0	-0.004622	-0.031230	0.782640
4008	0	1.123778	0.024747	-1.152941
4009	0	0.498438	-0.880209	-0.126473
4010	1	2.012576	1.452151	0.208426

변수마다 단위가 다르기 때문에 표준화를 통해 데이터 변환을 진행

3. Modeling 계획



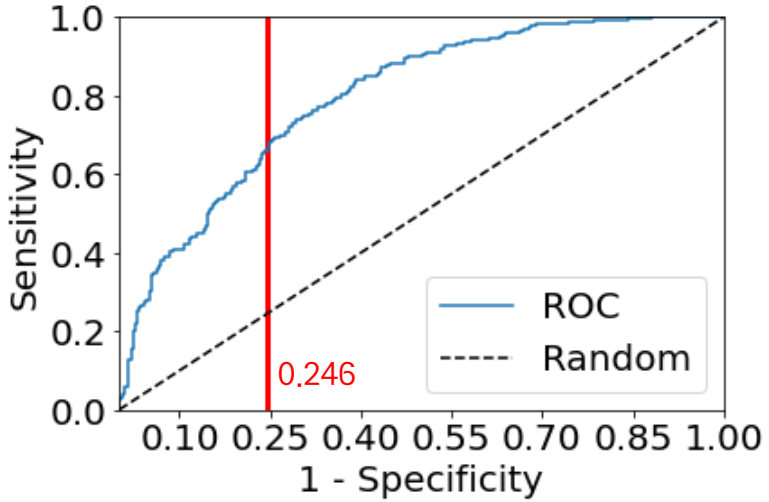
3. Modeling 로지스틱 회귀 모델

최적 하이퍼 파라미터 탐색

Parameter	search range		
C	1	~	10
max_iter	50	~	800

test set에 대한 모델 성능 : 0.697

ROC 곡선을 이용한 최적 임계값 산출



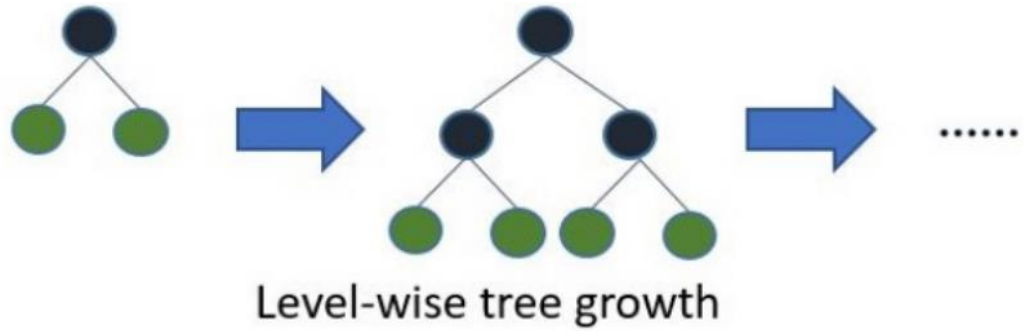
	Observed 0	Observed 1
Estimated 0	296	190
Estimated 1	51	266

Full Model					
변수	Intercept	Volume	Adult	Moisture	Density
Coef	-1.3870	0.3967	0.8245	0.9033	-0.3772

평가 지표					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.6999	0.8391	0.6091	0.5833	0.8530	0.6882

3. Modeling XGBoost 모델

경사하강법(Gradient Boosting) 알고리즘의
단점을 보완한 분류모델



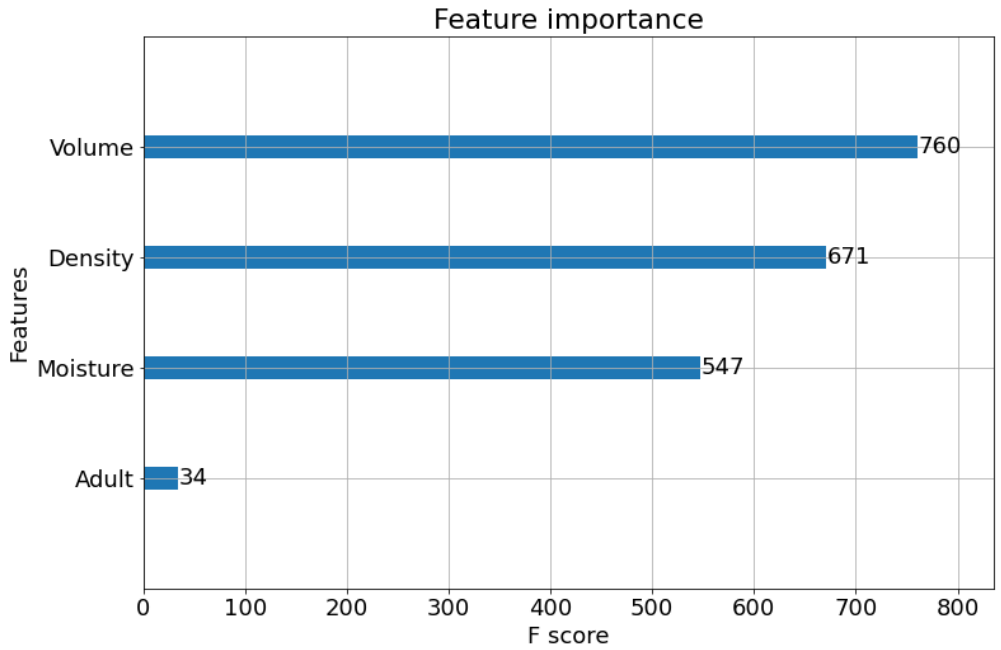
GBM에 비해 빠른속도, 규제를 통한 과적합 방지

3. Modeling XGBoost 모델

최적 하이퍼 파라미터 탐색 결과

Paramerter	Max_depth 10	N_estimator 100	Min_child_wiehgt 10
Cut-off value	0.307		
Accuracy	0.7491		

XGBoost 모형 변수 중요도



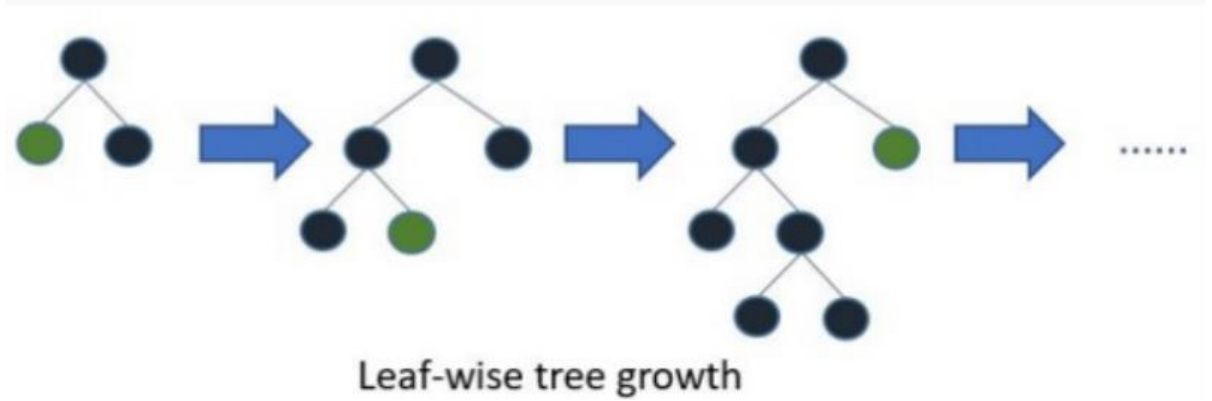
test set에 대한 성능

	Observed 0	Observed 1
Estimated 0	340	146
Estimated 1	81	236

평가 지표					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.6853	0.7445	0.6996	0.6178	0.8076	0.6853

3. Modeling LightGBM 모델

XGBoost 모델의 한계: 크게 개선되지 않은 속도
->처리 속도를 개선한 분류 모델



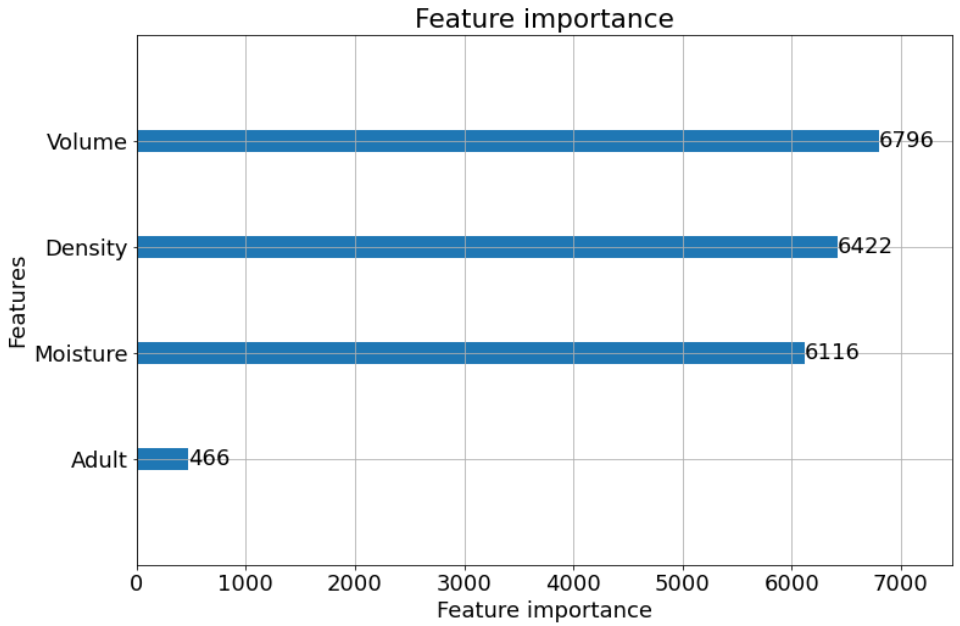
XGBoost보다 빠른 처리 속도, 과적합 가능성 상승
(너무 적은 데이터 사용시 과적합)

3. Modeling LightGBM 모델

최적 하이퍼 파라미터 탐색 결과

Paramerter	Max_depth 25	N_estimator 200	Num_leaves 10	Learning_rate 0.01
Cut-off value	0.343			
Accuracy	0.7504			

LgihtGBM 모형 변수 중요도



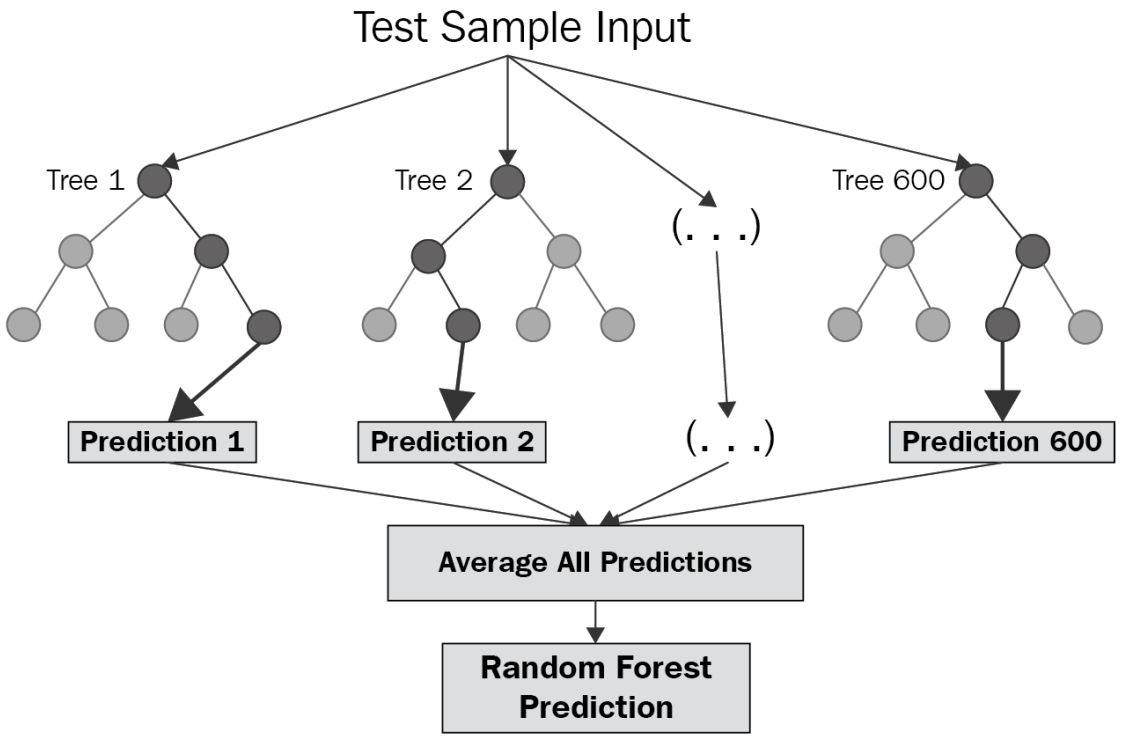
test set에 대한 성능

	Observed 0	Observed 1
Estimated 0	352	134
Estimated 1	97	220

평가 지표					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.7123	0.6940	0.7243	0.6215	0.7840	0.6557

3. Modeling 랜덤포레스트 모델

여러 숲(Forest)을 통해 자체적으로 일반화의 오류를 방지하는
분류 알고리즘



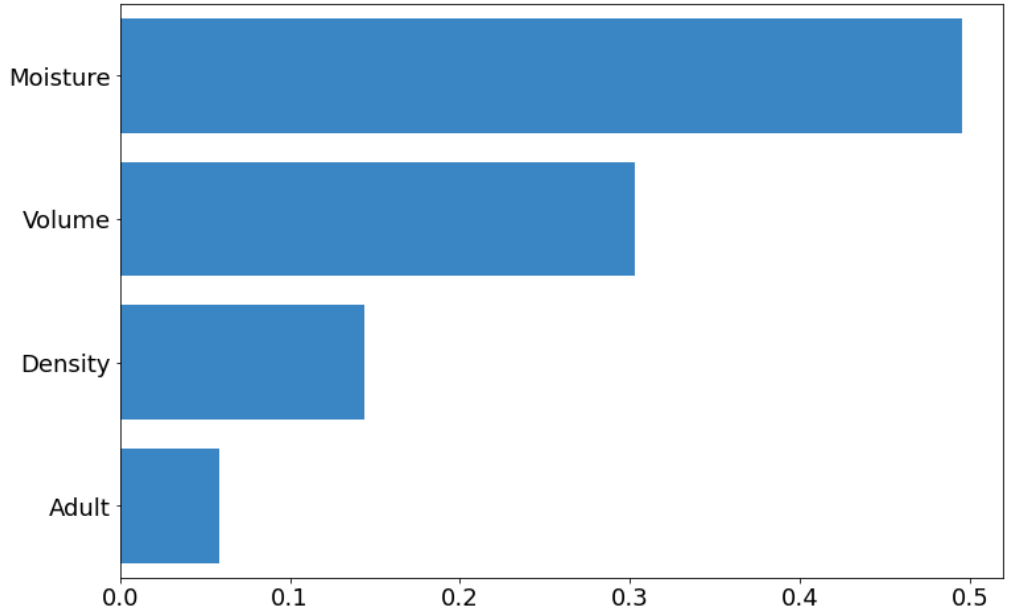
대부분의 경우 가장 좋고 안정된 성능 제공
레벨이 많은 속성에 편향적

3. Modeling 랜덤포레스트 모델

최적 하이퍼 파라미터 탐색 결과

Paramerter	Max_depth 10	N_estimator 100	Min_samples_leaf 10
Cut-off value	0.344		
Accuracy	0.7610		

RandomForest 모형 변수 중요도



test set에 대한 성능

	Observed 0	Observed 1
Estimated 0	336	150
Estimated 1	73	244

평가 지표					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.7223	0.7697	0.6914	0.6193	0.8215	0.6864

3. Modeling 모델 비교

Logistic Regression					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.6999	0.8391	0.6091	0.5833	0.8530	0.6882

XGBoost					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.6853	0.7445	0.6996	0.6178	0.8076	0.6853

LightGBM					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.7123	0.6940	0.7243	0.6215	0.7840	0.6557

RandomForest					
정확도	민감도	특이도	양성예측도	음성예측도	F1-score
0.7223	0.7697	0.6914	0.6193	0.8215	0.6864

모든 분류 모델 준수한 성능 확인

<양식장 입장>

성장 속도가 아직 빠른 전복을 판매하는 것
성장 속도가 느린 전복을 계속 키우는 것
= 손해

두 경우를 모두 고려했을 때 다른 지표보다
정확도가 가장 중요하단 결론 도출

4. 결과 해석 Insight

성장율이 하락세로 변한 전복의 특징

Volume  Adult 

무게 / 단위면적 

“최적의 전복 판매 타이밍을 잡는데 도움이 될 지표들”

4. 결과 해석 연구 한계 및 발전 방안

대부분의 변수가 사실상 한 지표를 따라 감
(설명력 있는 변수의 부족)

→ 변수변형 및 파생변수로 최대한 상쇄

물리적인 특성으로만 제한된 분석

→ 먹이와 서식환경에 대한 변수가 확보된다면
더욱 좋은 분석이 가능할 것으로 예상

감사합니다