강북삼성병원 갱년기 여성 건강 과제

데이터 정제 로그

성균관대 박지성, 이선민

0 Intro

정제 순서도

데이터 병합

- ☑ 병합 Codebook 형성
- ☑ 데이터 형식 통일

정렬 및 중복 행 제거

- ☑ 철회자 제외
- ☑ 정렬 기준 생성 및 형성

월경 주기 계산

- ☑ Cycle length
- ☑ Cycle length Difference

연구 데이터 셋 형성

- ☑ 연구 대상자 선정
- ☑ 대상자 구분 변수 생성

1

데이터 병합

1 데이터 병합

Raw data 정보

병합 대상 데이터

	조사 주기	조사 방법	활용 데이터
중년여성 건강 특화 설문조사	연 1회	비대면 조사 (문자)	환자 번호, 설문 조사일, 가장 최근 생리일, 마지막 이전 생리일
월경주기 조사	조사 년도 마다 상이	비대면 조사 (문자)	연구번호, 환자 번호, 생리시작일, 응답일, 월경 규칙성 여부, 월경 횟수
검진(문진)	매 2년 1회 이상	대면조사 (문진 및 검진)	환자번호, 검진일, 가장 최근 생리일의 년, 월, 일 마지막 이전 생리일의 년, 월, 일

참고 용 데이터

Master 파일

: 환자 별 통일된 생년월일, MP번호 및 월경 주기 추적 조사 문자 송부일자 및 답변일자 일부 존재

1 데이터 병합

병합 기준 Codebook : 총 11개의 변수로 구성

patient_id	연구 번호
enroll_id	MP 번호
birthdate	생년월일
mensdate	환자의 생리일
answer_date	응답 날짜 및 시간 : raw 데이터가 지니고 있던 정보 / 모두 정확한 년-월-일의 형식
answer_date_sub	추정 응답 날짜 : 응답일에 대한 일부 요소를 알 수 없는 경우 추정 가능한 정보까지 작성
mense_regul	환자가 직접 기입한 월경의 규칙성
data_from	원본 데이터 명
cycle	각 년도의 조사에서 환자가 본인의 생리일을 작성할 수 있는 칸의 개수
n	환자가 직접 작성한 본인의 생리 횟수
memo	비고

1 데이터 병합

데이터 형식 통일

중년 여성 설문조사, 검진 데이터

✓ 분석 목적에 맞는 필요 변수 만 추출한 데이터 셋 구성

설문조사 | ksw_mens_recent, ksw_mens_last_recent
검진 | kfchek4_11, kfchek4_12, kfchek4_13, kfchek5_11, kfchek5_12, kfchek5_13

✓ Master파일을 참고하여 각 patient_id에 해당하는 enroll_id, birthdate 부여

월경 주기 조사 데이터

✓ 2014~2020 년도 : 불필요한 변수 제거 및 데이터 형식 통합 과정 진행

✓ 2021 년도 : mensdate을 기준으로 answer_date_sub 추정 진행
long-form으로 전환 및 데이터 형식 통합 과정 진행

✓ 2022 년도 : 환자가 직접 기입한 응답 날짜 이전까지의 생리 횟수(n)에 관한 정보 존재
long-form으로 전환 및 데이터 형식 통합 과정 진행

2

정렬 및 중복 행 제거

2 정렬 및 중복 행 제거

☑ 앞선 병합 과정을 통해 환자 5,246명에 해당하는 457,133행의 dataset 형성

1) 철회자 제거

✓ master 파일을 기준으로 철회 이전까지의 데이터 활용에 대한 동의 여부에 해당하는

✓ [off_dataagree] 의 값이 1(비동의)에 해당하는 환자들에 대한 모든 정보 제거

229명의 patient_id에 해당하는 7,375개의 행 제외

2) [mensdate_trans] 파생 변수 생성

✓ 생리일 (mensdate)에 관하여 정확한 날짜가 기입되어 있지 않은 경우 작성된

✓ 형식이 데이터 마다 다양함

✓ 원본 데이터 유지를 위해 [mensdate] 값을 그대로 두고, [mensdate_trans] 를

이용해 각각의 정보 유지

2 정렬 및 중복 행 제거

2) [mensdate_trans] 파생 변수 생성

유형	mensdate	mensdate_trans
mensdate가 무응답인지 또는 결측치에 해당하는 지 구분이 어려운 경우	none / None / 기억안남 / NA / -1 / -3 / 없음 / 9999-99-99 / 무월경	“99”로 기입
[mensdate] 값이 날짜 형태로 기입되어 있으나 정확한 추정이 불가능한 경우	1918-05-16 / 617-02-99	
[mensdate]값이 일부만 존재하나 추정이 가능한 경우	2019년 8월 / 2020년 6월 24-26일	날짜의 요소가 일부 누락된 경우 해당 요소를 “99”로 기입

2 정렬 및 중복 행 제거

정렬

☑ 환자 별 월경 주기를 산출하기 위해 선행 되어야 하는 작업

1차 정렬 [patient_id], [mensdate], [mense_regul] 순으로 정렬

동일한 환자, 생리일에 대해서 가장 정보가 많은 행이 제일 위에 정렬되도록 하기 위함

2차 정렬

조건문 순서	data_from	mensdate_trans	answer_date	**정렬 대상**
1	2021년도 월경주기 조사	99	NA	**answer_date_sub**
2	그 외	99	Not NA	**answer_date**
3	그 외	그 외	그 외	**mensdate_trans**

2 정렬 및 중복 행 제거

중복 행 제거

1차 제거 [data_from] 변수를 기준으로 제거

✓ 한 행을 이루는 모든 변수가 완벽하게 일치하는 경우 제거

✓ 각 원본 데이터가 지니는 속성과 결측 치의 특징을 고려하여 제거

e.g) 변수 n(환자가 직접 기입한 생리 횟수)을 기준으로 환자 별 해당 횟수에 해당하는 행을 남기고, 만약 변수 n보다 작성된 생리일의 수가 더 많다면 데이터 유지를 위해 작성된 생리일도 우선 보존하는 방식으로 진행 **(2022년도 월경주기 조사 데이터에 해당)**

2차 제거 데이터의 유형 별 제거

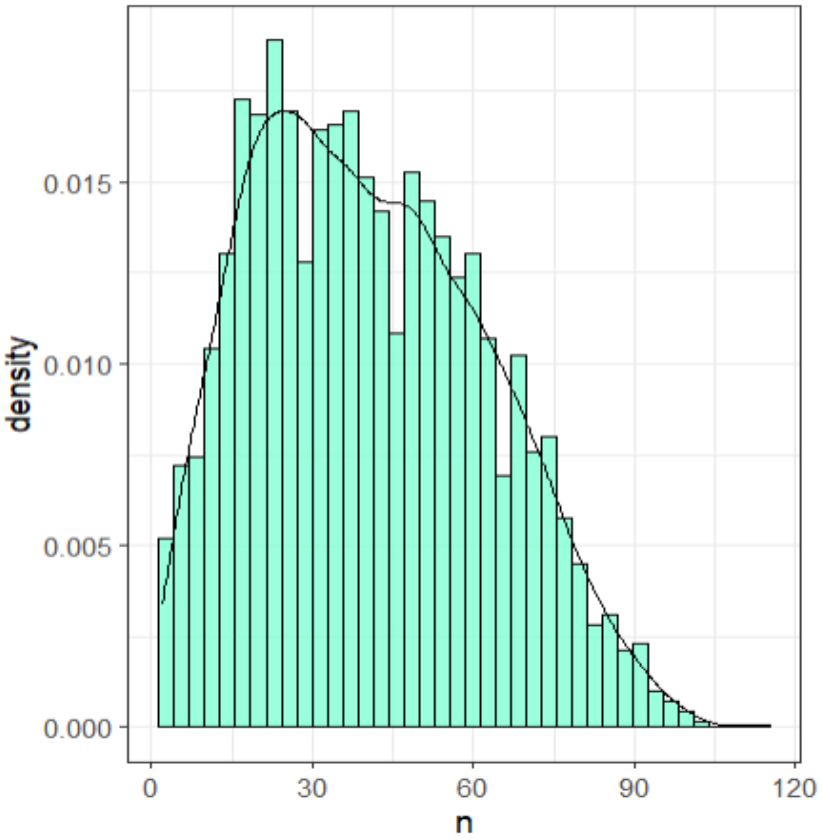
✓ 온전한 값으로 중복되어 존재하는 [mensdate] 행에 대해, [mensdate]과 가장 가까운 [answer_date]에 해당하는 행만 남기고 나머지는 제거

✓ 최종적으로, 각 데이터 유형에 대하여 동일한 patient_id, mensdate에 대해 1개의 행만 남음

2 정렬 및 중복 행 제거

시각화 : 정렬 및 중복 행 제거가 완료된 전체 병합 데이터 사용

환자 별 생리 기입 횟수 분포

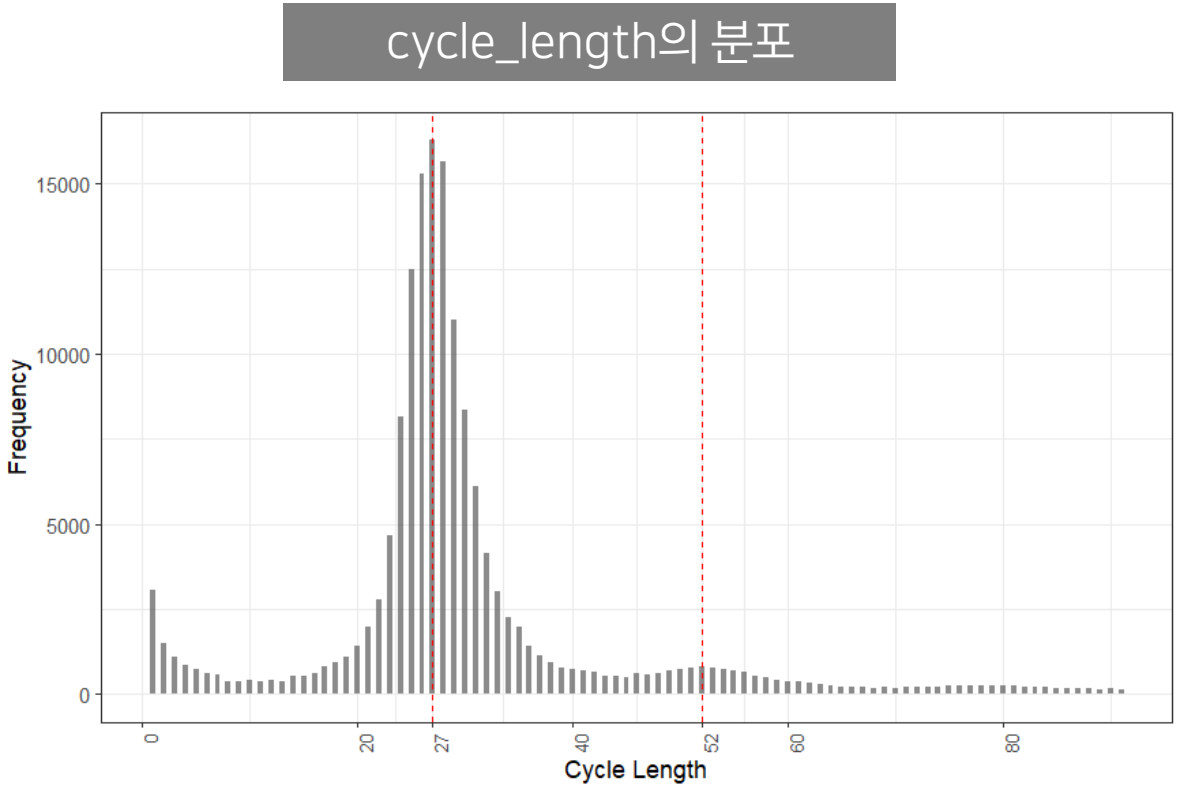


✓ 20~30회 부근에서
생리일을 기입한
개수의 빈도가 가장 높음

✓ 대부분의 patient_id에 대하여
생리일 기입 횟수가
10회 이상 기록됨을 확인

2 정렬 및 중복 행 제거

시각화

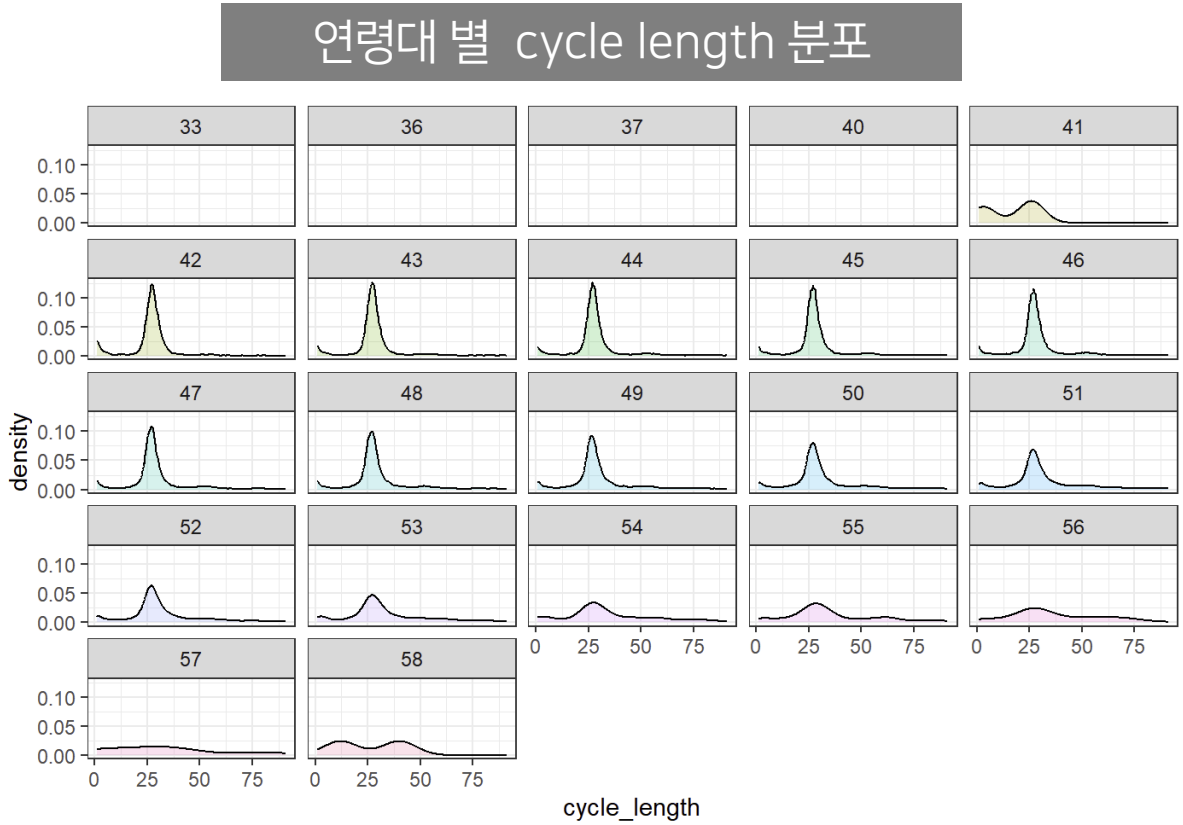


✓ cycle_length가 27인 경우의 빈도가 가장 높았음

전체 cycle length 의 분포 중 중위값인 28과 유사함

2 정렬 및 중복 행 제거

시각화

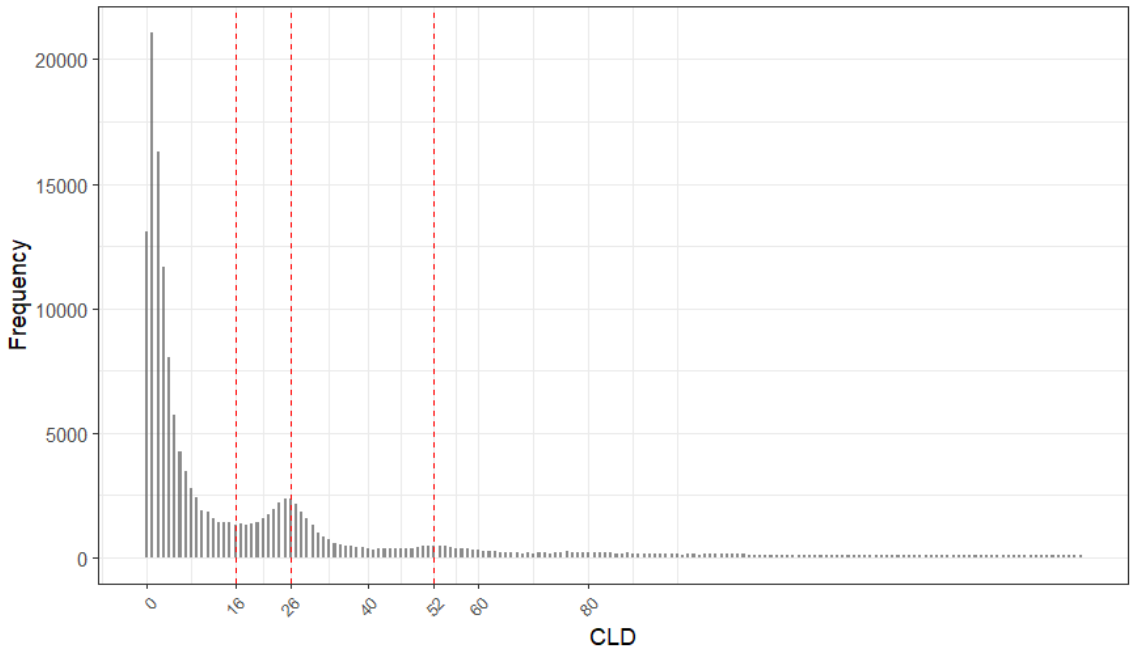


✓ 연령이 42세 이상에서 53세까지는 전체 cycle length 의 분포와 유사하나,
그 이후의 연령대에서는 고르게 분포되어 있는 것을 확인할 수 있음

2 정렬 및 중복 행 제거

시각화

CLD 분포

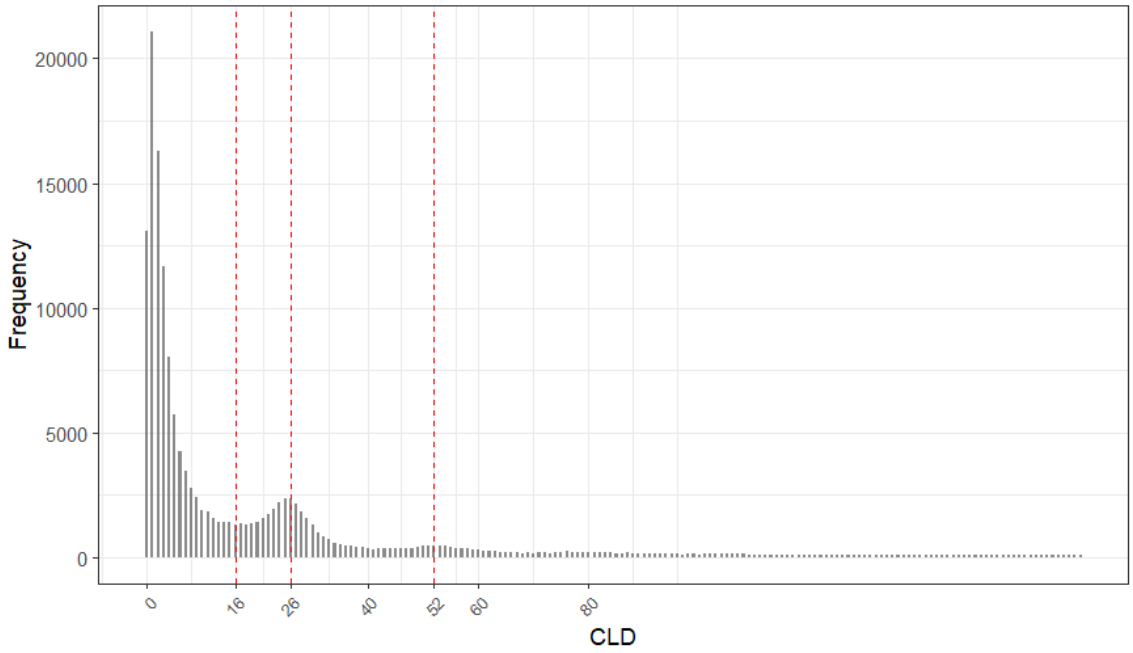


✓ 전반적으로 CLD값이 커짐에 따라 빈도가 줄어드는 경향을 보임

2 정렬 및 중복 행 제거

시각화

CLD 분포



✓ CLD가 26, 52일 경우에 대해서 소폭 상승하는 경향을 보임

폐경 이행 단계를 겪는 연령 대의 데이터이기에 소폭 상승의 요인이 artifact 혹은 본래 데이터의 특성인지 분간하기 어려움

2 정렬 및 중복 행 제거

시각화

전체 데이터의 특성

- ✓ 대부분의 환자에 대해 기입된 생리 횟수가 20회 이상
- ✓ Cycle Length 가 27, 28의 값을 가지는 비율이 가장 높았으며, 해당 값은 규칙적인 생리 주기로서 의미를 가짐
- ✓ 연령대별 Cycle Length 의 분포를 보았을 때, 연령이 증가할수록 Cycle Length 의 불규칙성이 두드러지는 특성을 보임

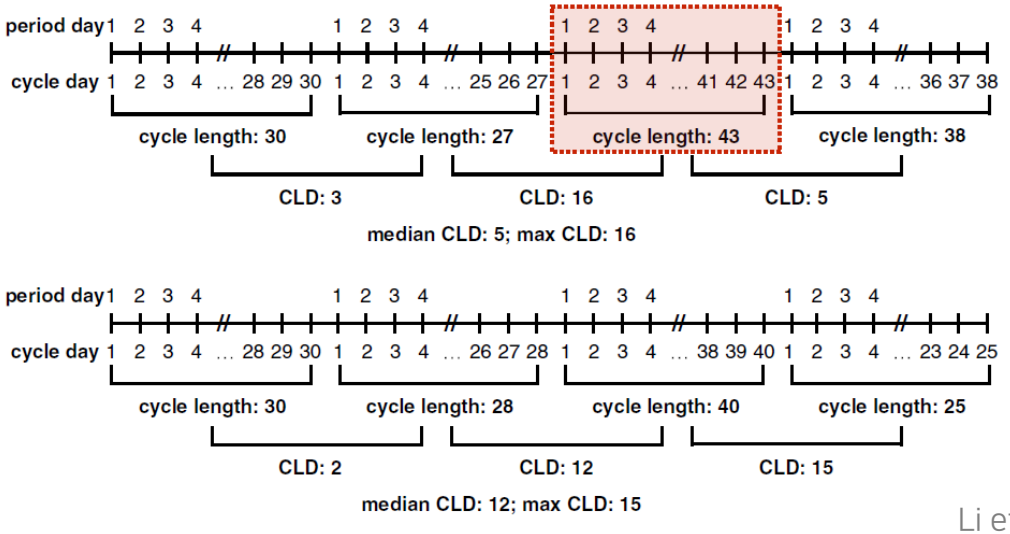
병합된 전체 데이터의 시각화를 통해 병합된 데이터가
중년 여성의 월경에 대한 주기적 특성을 잘 반영하고 있다고 판단함

3

월경 주기 계산

3 월경 주기 계산

월경 주기 계산



Li et al. (2020)

- 환자 별 연속된 생리 일의 차이를 계산하여 Cycle Length 산출
- 이후, 연속된 CL의 차이를 바탕으로 Cycle Length Difference(CLD) 산출

4

연구 데이터 셋 형성

4 연구 데이터 셋 형성

연구 대상자 선정

연구 대상자 선정 목적

✓ 앞서 병합한 데이터 셋의 정확성 확보 및 신뢰성 검증하기 위한 최종 단계

✓ 환자 별 검진 데이터를 바탕으로 신뢰도가 가장 높다고 판단되는 환자들을 추출하여 해당 환자들에 대한 CL 및 CLD의 패턴을 확인

- 위의 기준들로 추출된 표본이 전체 데이터를 대표할 수 있는 지의 여부를 판단하는 시각화 및 통계적 검증 과정을 함께 진행

4 연구 데이터 셋 형성

연구 대상자 선정 기준

	삭제 내용	삭제 단위
검진 기준	수유 / 임신 / 암 진단력 / 자궁적출술 또는 난소절제술 / 호르몬 치료 / 피임	환자
생리 기입 횟수	생리 기입 횟수가 10회 미만	환자
생리의 규칙성(1)	검진 데이터 기준 [enroll_date] 시점에 생리 규칙성을 “규칙적이다 " 라고 응답한 경우	환자
생리의 규칙성(2)	환자 별 가장 이전 세 개의 [CLD] 값이 모두 3 이하인 경우	환자
나이(1)	42세 미만인 경우	행
나이(2)	42세 미만이고 [enroll_date]를 기준으로 이전 92일 이내에 [mensdate]가 작성되지 않은 경우	행

4 연구 데이터 셋 형성

최종 연구 데이터 셋 형성

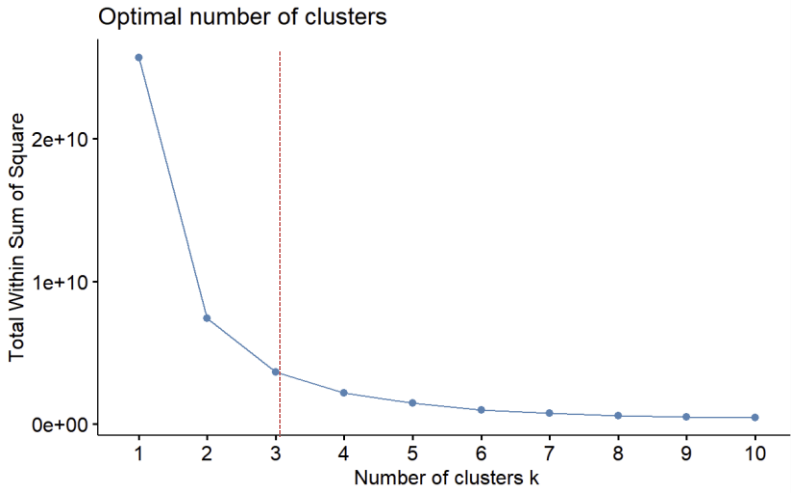
데이터 명	구성 조건	데이터 행
Data Set 2	연구 대상자 선정	491명의 환자에 대한 20,598행
Data Set 2-1	4SD 기준으로 이상치 제거 후, 연구 대상자 선정	761명의 환자에 대한 31,159행
Data Set 3	환자의 [meno_stage]를 고려하여 폐경 이행기를 겪지 않은 연구 대상자만 선정	410명의 환자에 대한 8,115행
Data Set 3-1	4SD 기준으로 이상치 제거 후, 폐경 이행기를 겪지 않은 연구 대상자만 선정	636명의 환자에 대한 11,874행

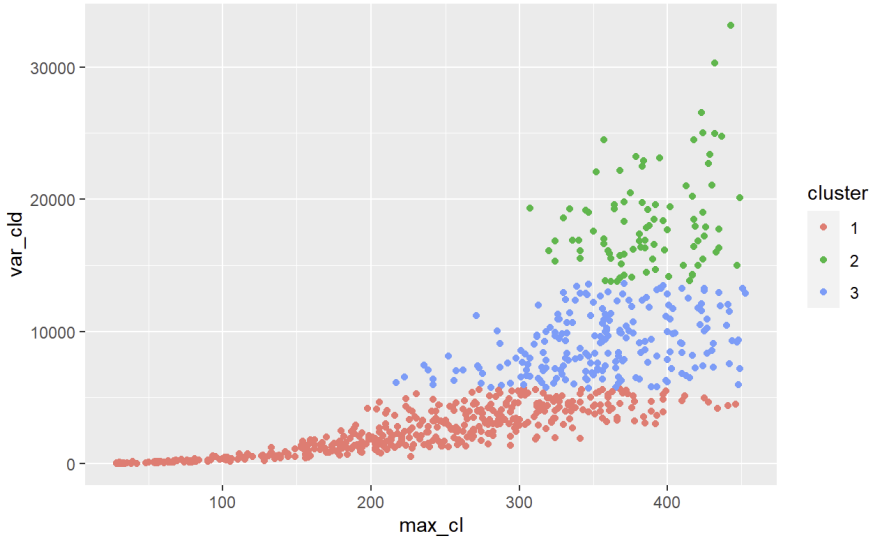
4 연구 데이터 셋 형성

연구 대상자의 특징 변수 형성

☑ 형성한 최종 데이터 셋 중에서 최종적으로 유지된 patient_id과 그에 대한 정보가 가장 많았던 dataset2_1에 대하여 우선적으로 연구 대상자 별 특징 변수 형성 진행

cluster

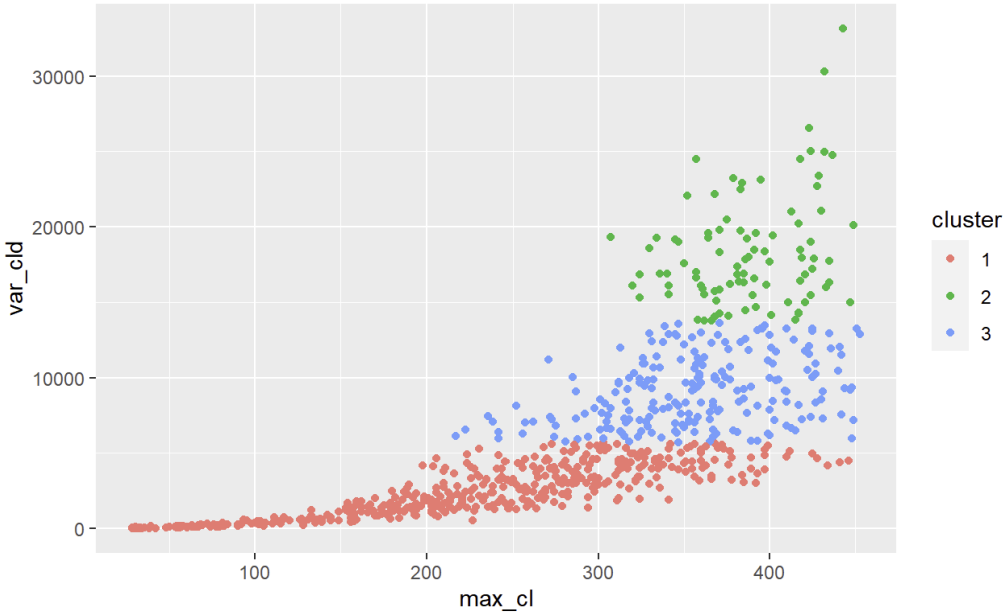




patient_id별로 생리주기(cycle_length)의 최대값과 CLD의 분산에 대한 scatter plot

4 연구 데이터 셋 형성

연구 대상자의 특징 변수 형성



Cluster 별 특징

Cluster 1

비교적 유사한 cycle_length가 반복적으로 발생하는 그룹

Cluster 3

3개의 cluster 중 cycle_length의 최대값 및 그에 대한 변동성이 중앙에 위치한 경우

Cluster 2

모든 cycle_legth에 대한 최대값이 300을 초과하며, 생리주기에 대한 변동성도 가장 큼

4 연구 데이터 셋 형성

연구 대상자의 특징 변수 형성

initial_stage

연구 대상자 별로 처음 연구 등록 시점(enroll_date)에
어떠한 폐경 이행 단계(meno_stage)를 겪고 있었는지에 관한 변수 생성

✓ 환자들 마다 데이터 상 등록 시점(나이)가 상이하기에
연구 시작시점에 겪고 있는 폐경 이행 단계도 상이함을 파악

✓ 이후 연구의 방향성에 따라 연구 참여 시점에서의 폐경 이행 단계에 따른
연구 대상자를 용이하게 구분할 수 있도록 하기 위해 해당 변수를 생성

premenopause	Early transition	Late transition	menopause	NA
656명	86명	10명	1명	8명

THANK YOU