2021 데이터 청년 캠퍼스

영유아 발달 상황 분석 알고리즘 개발

프릭스헬스케어 팀

Kdata 서울여자대학교 닥터 아이 Dr.i

INDEX

0. 팀 소개

강혜빈

김은영

엄소현

조혜리

홍예지

CONTENTS

1

데이터 분석 연구 배경

‘닥터아이’ 서비스

목표

2

데이터 분석 진행 과정

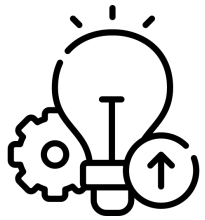
데이터 전처리

데이터 분석

알고리즘 개발

최종결과

3



적용과 기대효과

기존 서비스와의 차별성

활용방안과 기대효과

‘닥터아이’ 서비스

“닥터아이”는 부모가 아이 발달을 직접 체크하고 비교분석할 수 있는 앱

월별 발달 및 비교 분석

- 월별로 아이가 할 수 있는 행동을 알려준다.
- 아이가 처음 행동한 날을 퀘스트(마일스톤) 형식으로 기록한다.
- 아이가 또래에 비해 발달이 빠르고 느린 분야를 알려준다.

‘닥터아이’ 서비스

1

영유아의 발달 지연은 시기 적절한 치료가
중요하지만 **가정에서 영유아의 발달과 건강 관리를 전문적으로 할 수 있는 방안이 마땅치 않아**
많은 부모들이 어려움을 겪고 있음

2

부모들은 아이가 시기별로 잘 자라고 있는지
늘 의문을 가지고 있으나, **정기검진(영유아 검진) 외에는 확인할 수 있는 방법이 없음**

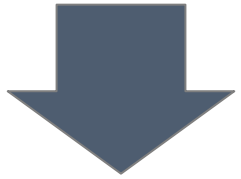
3

짧은 상담 시간, 부모의 시간부족으로 인한 문제 발생
및 검진 정보 부족, 검진 항목 과다, 형식적인
검진이라는 의견이 많음. 건강검진 필요성에 대한
인식이 저하되는 **영유아 검진 자체의 문제점도 제기됨**

목표

영유아 발달 상황

분석 알고리즘 개발



마일스톤별 도달 시점의 평균, 해당 user의 상위 %

ex:) 평균적으로 아이들이 걷게 되는 시점은 n일이다 or n일에 걸으면 상위 n%이다

CONTENTS

데이터 분석 연구 진행 과정

데이터 전처리

데이터 클렌징

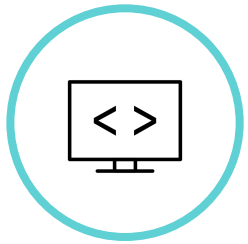
데이터 매핑

데이터 분석

생존분석

성별분석

알고리즘 개발



마일스톤별 생후일자 평균

마일스톤별 user의 상위 %

업데이트

분석 데이터 RAW DATA

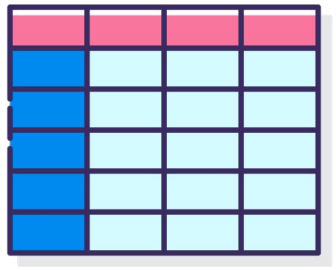
	A	B	C	D	E	F
1	person	milestone	grade	done_day_after_bi	created_at	updated_at
2	2677	1	4	110	2021-06-26 7:45:36	2021-06-26 7:45:36
3	3001	1	3	130	2021-06-28 1:14:51	2021-06-28 1:14:51
4	2166	1	4	140	2021-06-22 14:23:53	2021-06-22 14:23:53
5	2661	1	4	114	2021-06-26 7:16:52	2021-07-02 11:55:18
6	2948	1	4	156	2021-06-27 17:06:10	2021-07-04 10:35:32
7	2997	1	4	175	2021-06-28 0:40:15	2021-06-28 0:40:15
8	1987	1	3	83	2021-06-20 23:23:01	2021-06-20 23:23:01
9	2169	1	4	192	2021-06-22 14:21:41	2021-06-22 14:21:41
10	2553	1	4	146	2021-06-25 17:56:38	2021-06-25 17:56:38
11	2670	1	4	199	2021-06-26 6:19:50	2021-06-26 6:19:50
12	2741	1	4	157	2021-06-26 16:39:19	2021-06-26 16:39:19
13	2947	1	4	105	2021-06-27 16:53:00	2021-06-27 16:53:00
14	393	1	4	24	2021-06-27 19:23:21	2021-06-27 19:23:21
15	2995	1	4	103	2021-06-27 23:58:38	2021-06-27 23:58:38
16	1932	1	4	198	2021-06-20 1:54:49	2021-06-20 1:54:49
17	1983	1	4	145	2021-06-20 22:36:23	2021-06-20 22:36:23
18	2046	1	2	71	2021-06-21 13:14:57	2021-06-21 13:14:57
19	2160	1	4	173	2021-06-22 13:27:52	2021-06-22 13:27:52
20	2192	1	3	134	2021-06-22 18:31:52	2021-06-22 18:31:52
21	2537	1	4	119	2021-06-25 14:01:33	2021-07-15 0:26:14
22	2626	1	4	198	2021-06-25 23:47:05	2021-06-25 23:47:05

분석 데이터 RAW DATA

person	milestone	grade	done_day_after_birth	created_at	updated_at
2677	1	4	110	2021-06-26 07:45:36	2021-06-26 07:45:36
2661	1	4	114	2021-06-26 07:16:52	2021-07-02 11:55:18

- person : 사용자 고유 ID
- milestone : 질문 넘버
- grade : 응답 번호 (단, 1로 응답한 데이터는 기록되지 않음)
- done_day_after_birth : 사용자가 입력한 유아의 생후일수
- created_at : 마일스톤 입력 시작 일
- updated_at : 최종 마일스톤 입력 일

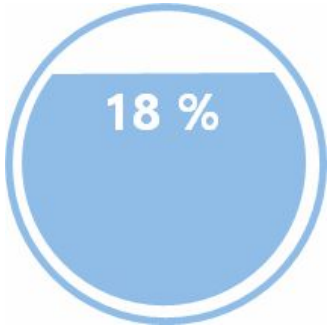
데이터 전처리 데이터 클렌징



원본 데이터

2021-04-18 0:00:00 ~ 2021-07-27 18:35:53

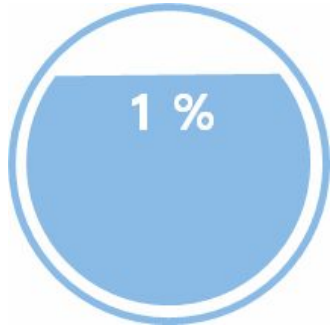
raw data 총 301,496개



중복 제거

중복 데이터 제거

▶ 244,568 개



0시 제거

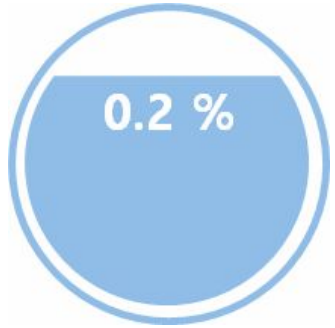
0:00:00:00 시간 제거

▸ 241,298 개

분석 유효 범위 설정

응답 분포에 따른 유효 범위 설정

▶ 240,760 개



Outlier 제거

IQR 방식을 사용하여 이상치 제거

▶ 239,988 개

데이터 전처리 데이터 클렌징

중복 데이터 앱 오류 혹은 개발팀 test 데이터

person	milestone	grade	done_day_after_birth	created_at	updated_at
367	1	2	97	2021-06-06 3:55:19	2021-06-06 3:55:19
367	1	2	97	2021-06-06 3:55:19	2021-06-06 3:55:19

created_at과 updated_at 시간이 모두 0:00:00인 데이터 개발팀 test 데이터

339	558	3	1050	2021-06-05 0:00:00	2021-06-05 0:00:00

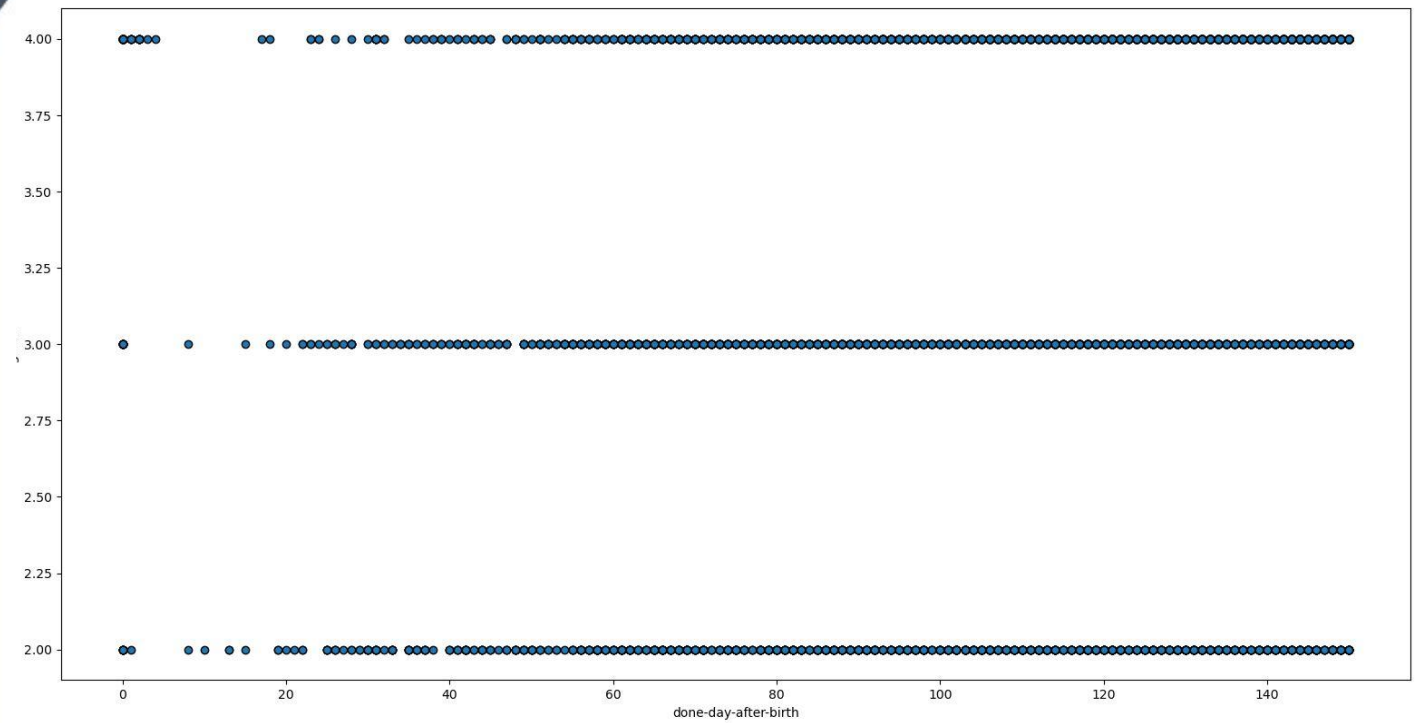
분석 비유효 범위 데이터 앱 출시(2021년 06월 03일) 이전에 기록된 데이터 : 개발팀 test 데이터

345	1	2	105	2021-05-14 12:57:48	2021-05-14 12:57:48

데이터 전처리 데이터 클렌징

분석 유효 범위 데이터 생후 21일 이전 응답 데이터

Scatter



전체 마일스톤에 대한 생후 1일부터 150일까지의 응답 분포도

Check

"마일스톤별 응답 분포"

마일스톤별 응답 데이터의
done_day_after_birth 분포를 확인해보니
20일 이후의 데이터가 연속적으로 존재

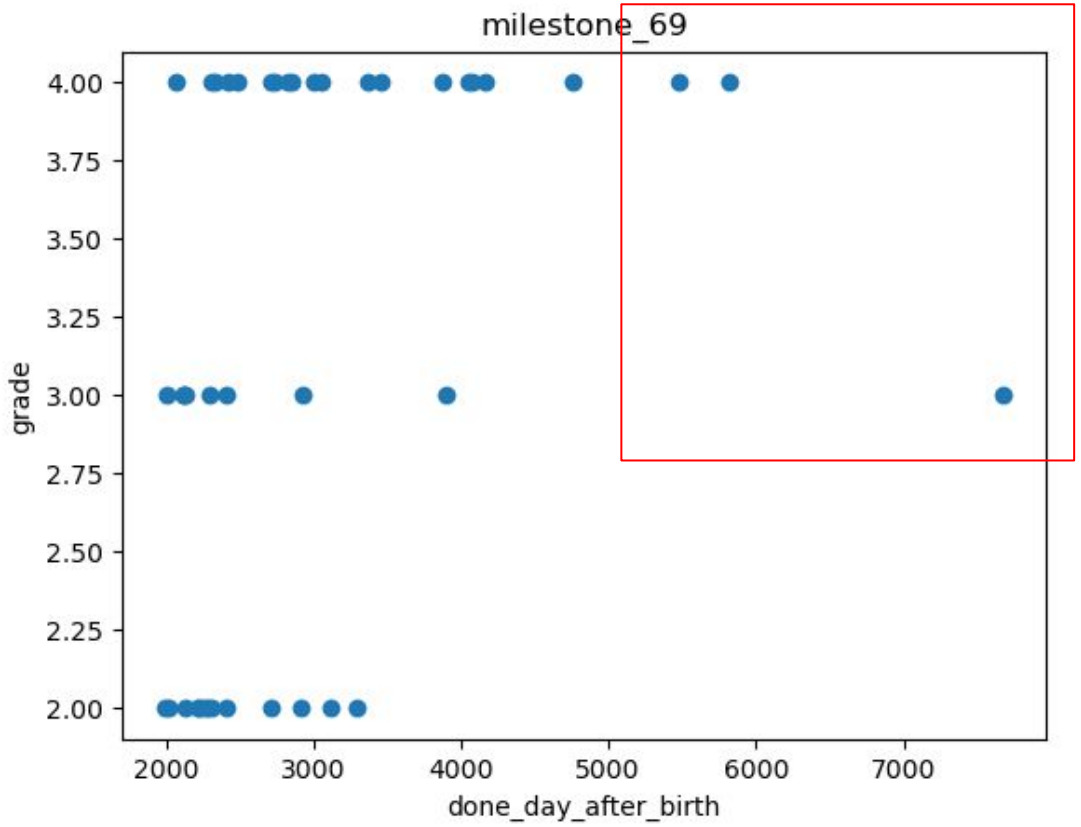
Apply

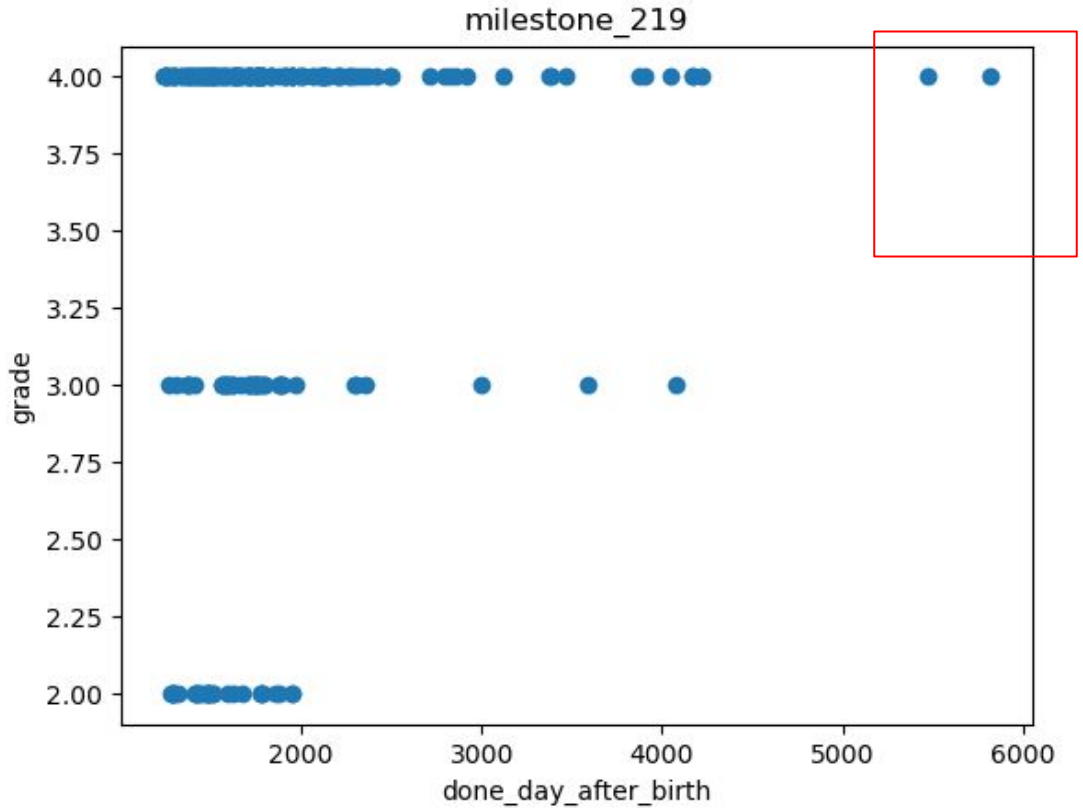
"20일 이후의 응답 데이터 사용"

- 생후 21일 이전 응답 데이터 非균일
- 생후 21일 이전 응답 데이터가 360개 밖에 되지 않음
- 응답에 해당하는 마일스톤이 생후 20일 이전의 아이가 할 수 있는 행동이라고 보기 어려움

데이터 전처리 데이터 클렌징

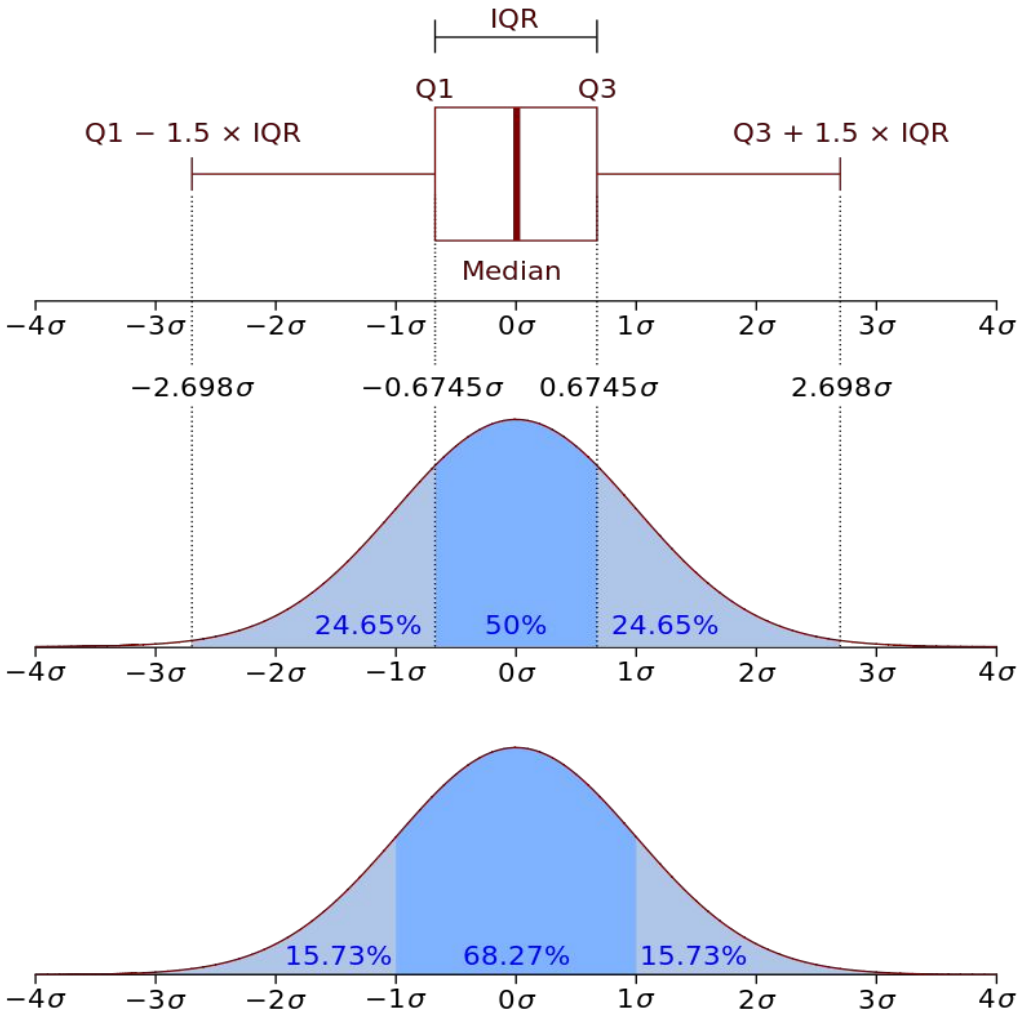
Outlier 정상범위를 벗어나는 데이터





마일스톤 넘버별 scatter를 찍어 보았을 때,
자료분석의 결과를 왜곡시키거나 적절성을 위협하는 이상치(outlier) 발생

데이터 전처리 데이터 클렌징



IQR(Inter Qunatile Range)

중앙에 위치한 중앙값의 좌우로부터 동일한 백분율을 가진
두 점간의 거리에 의해 알아보는 방법

자료의 극단적인 값들에 의한 영향을 덜 받는 장점

IQR = Q3(75%) − Q1(25%)

Q3+(IQR*1.5) → 최대값

Q1+(IQR*1.5) → 최소값

따라서 최대값(Q3), 최소값(Q1)보다 크거나 작은 값인 이상치 제거

데이터 전처리 결과

301,496 ▶ 239,988

약 20 % 제거

데이터 전처리 데이터 매핑

K-DST 기준 닥터아이 앱 내 milestone질문 : 총 800개

contents
가족 등 친숙한 사람을 보면 다가가려고 해요.
딸랑이를 손 가까이 주면 잡아요.
누워 있다가 혼자 앉을 수 있어요.
"엄마" 또는 "아빠"와 비슷한 소리를 내요. (의미없이 내는 소리도 포함)

unique milestone : 총 281개

contents	unique_milestone
가족 등 친숙한 사람을 보면 다가가려고 해요.	4
딸랑이를 손 가까이 주면 잡아요.	66
누워 있다가 혼자 앉을 수 있어요.	130
"엄마" 또는 "아빠"와 비슷한 소리를 내요. (의미없이 내는 소리도 포함)	173

중복되는 질문(800)들을 unique_number(281)로 매핑

gender 정보 매핑

person_id	gender
2677	F
2661	M

기존 데이터의 person과 person_id 기준으로 성별 정보 매핑

- gender 정보를 통해 남자아이, 여자아이에 따른 개별 분석 가능

데이터 전처리 최종 데이터

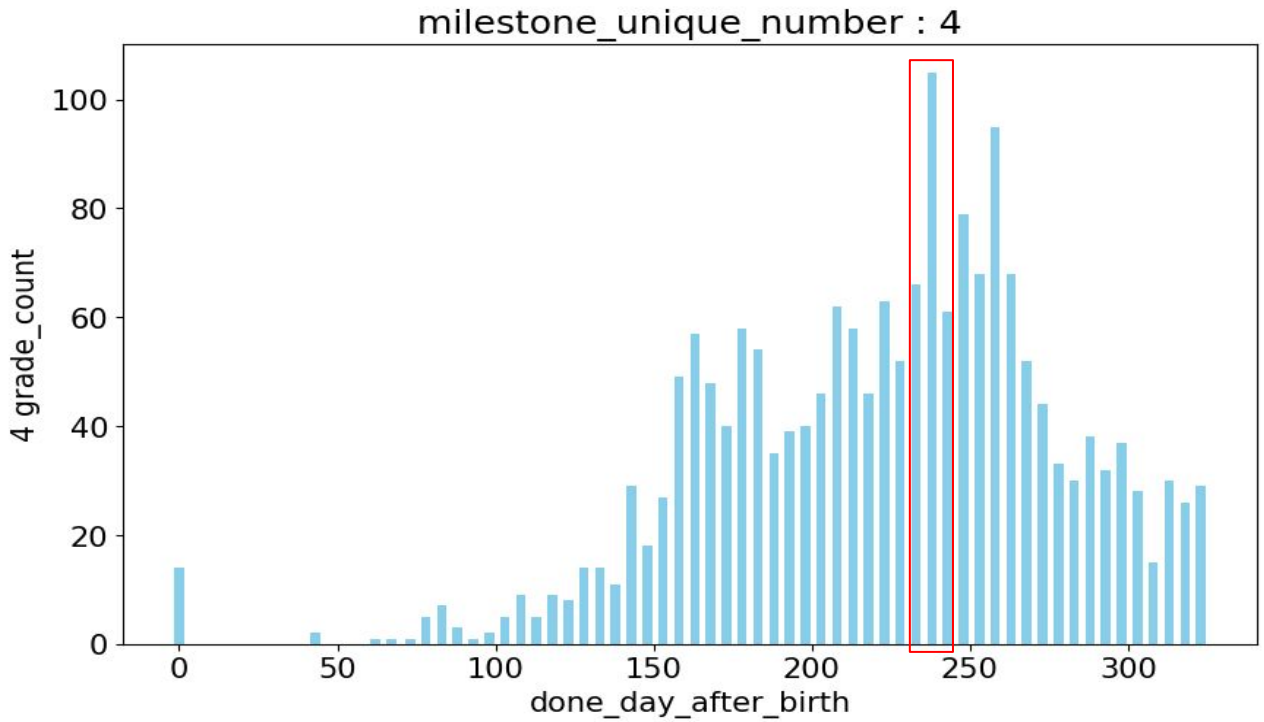
person	milestone	grade	done_day_after_birth	created_at	updated_at	unique_milestone	gender
2677	1	4	110	2021-06-26 07:45:36	2021-06-26 07:45:36	6	F
2661	1	4	114	2021-06-26 07:16:52	2021-07-02 11:55:18	8	M

- person : 사용자 고유 ID
- milestone : 질문 넘버
- grade : 응답 번호 (단, 1로 응답한 데이터는 기록되지 않음)
- done_day_after_birth : 사용자가 입력한 유아의 생후일수
- create_at : 마일스톤 입력 시작 일
- updated_at : 최종 마일스톤 입력 일

unique_milestone : 마일스톤 고유 넘버

gender : 성별 정보

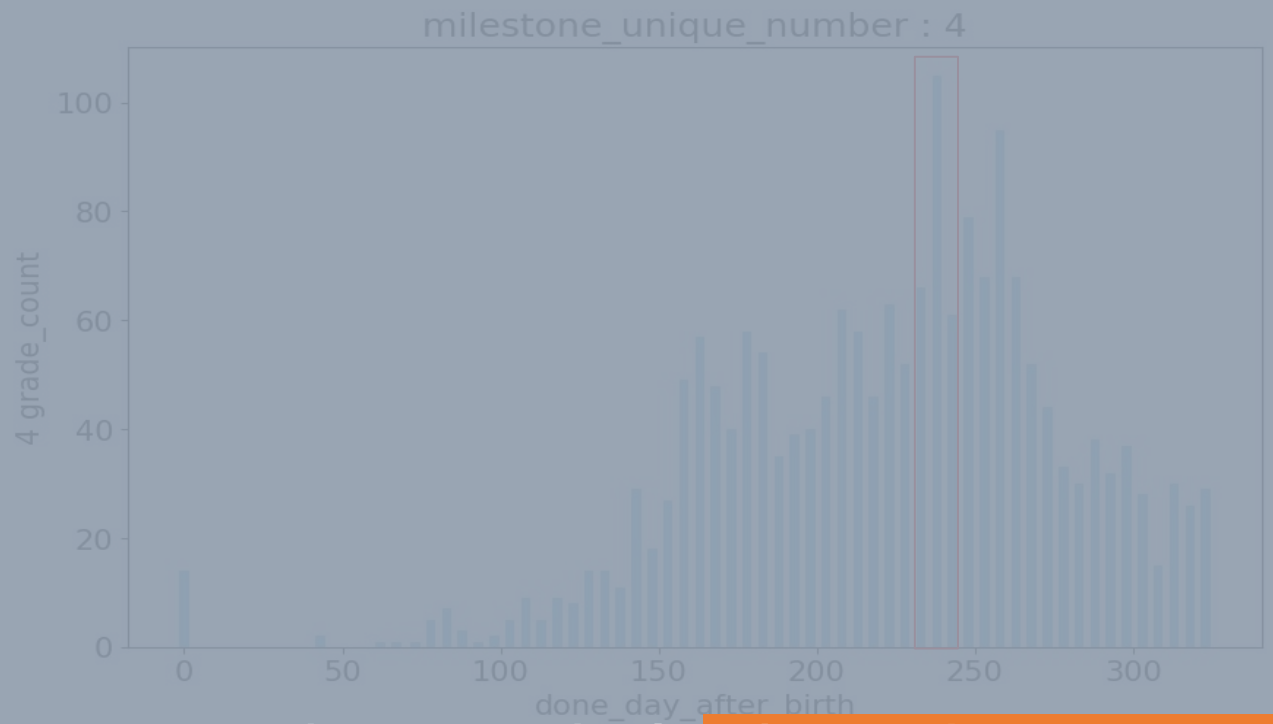
데이터 분석 시행착오



done_day_after_birth 5일씩 그룹화 후 4번 응답이 가장 많은 구간 추출

구간에서 4번 응답이 가장 많은 일자 = 평균적으로 가능한 시점

데이터 분석 시행착오



user가 입력했을 시점은 아이가 이미 마일스톤을 도달한 후,

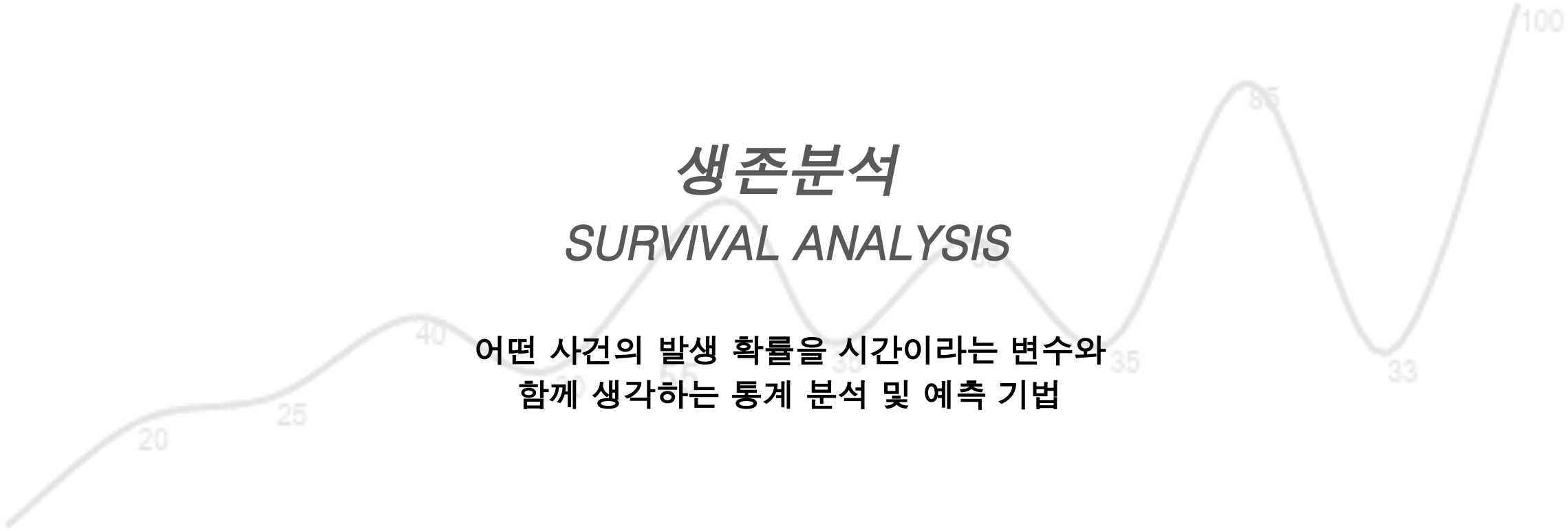
따라서 이 통계 분석은 옳지 않다.

구간에서 4번 응답이 가장 많은 일자 = 평균적으로 가능한 시점

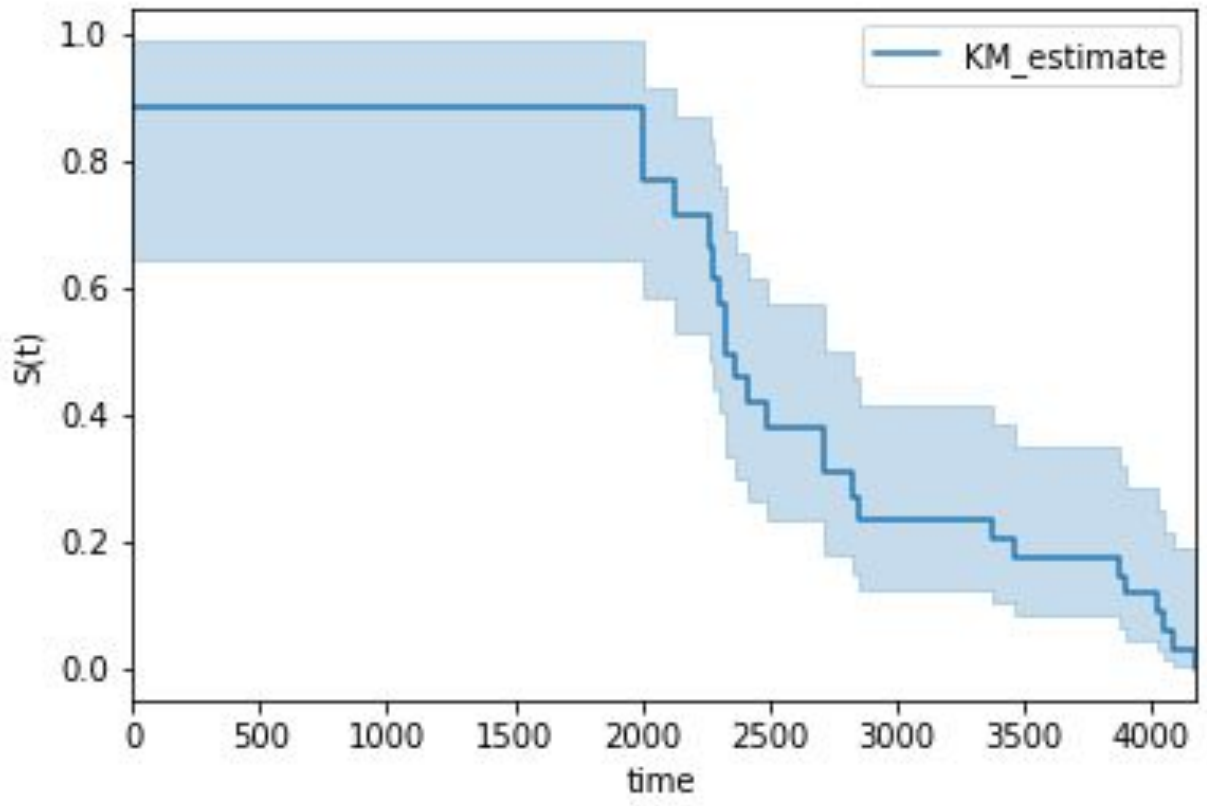
생존분석

SURVIVAL ANALYSIS

어떤 사건의 발생 확률을 시간이라는 변수와
함께 생각하는 통계 분석 및 예측 기법



데이터 분석 KAPLAN-MEIER



KAPLAN-MEIER

생존분석에서 사용되는 통계 기법

데이터의 관찰 시간에 따라서 사건이 발생한
시점에서의 사건 발생률을 계산하는 방법

$S(t_i)$의 생존 시간 확률

$$S(t_i) = S(t_{i-1}) * (1 - \frac{d_i}{n_i})$$

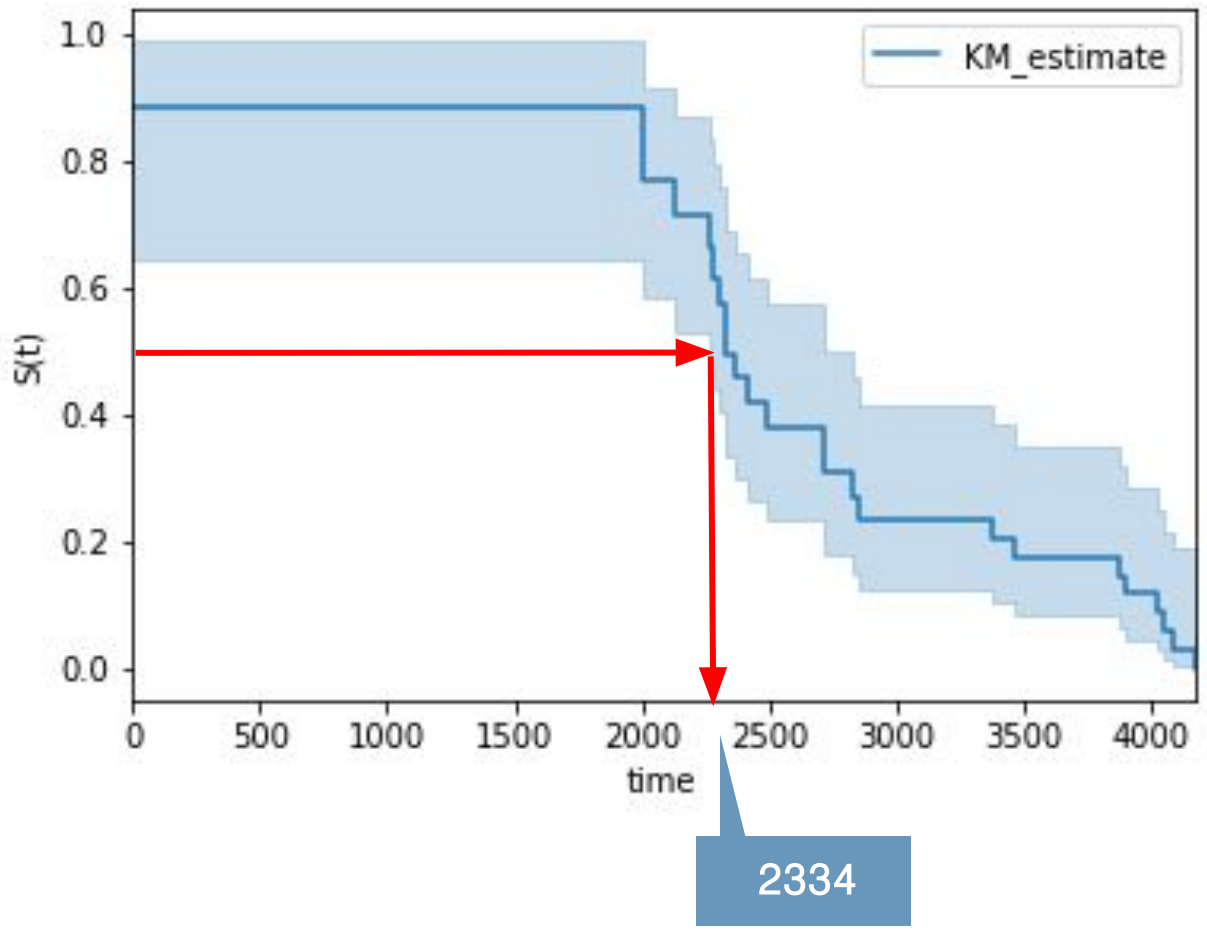
$S(t_i)$ = The probability of being alive at time t_i

n_i = The number of subjects alive just before time t_i

d_i = The number of events at time t_i

$S(0) = 1$, $t_0 = 0$

데이터 분석 KAPLAN-MEIER



KAPLAN-MEIER

생존분석에서 사용되는 통계 기법

데이터의 관찰 시간에 따라서 사건이 발생한
시점에서의 사건 발생률을 계산하는 방법

마일스톤 별 평균 일수

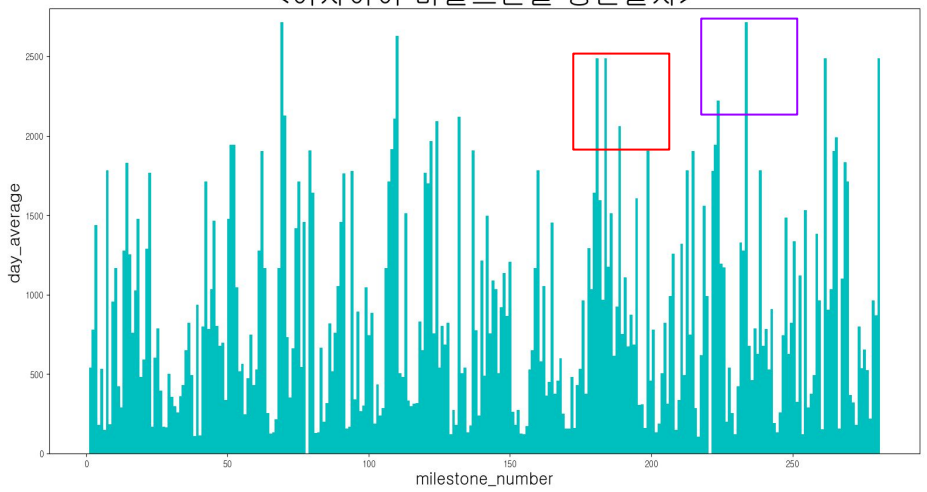
: 평균적으로 아이의 50%가 도달한 생후 일수

kmf.median_survival_time_

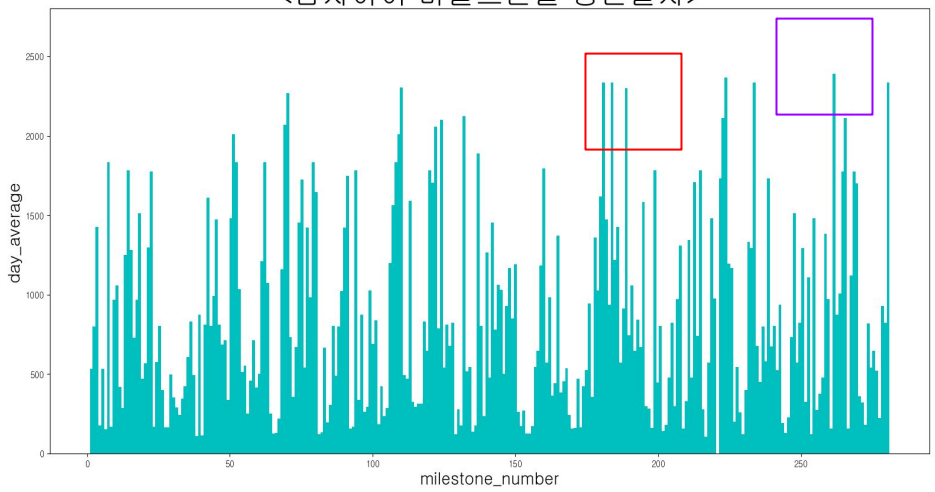
데이터 분석 성별 비교

※ 성별에 따른 마일스톤(행동) 가능일자 비교

<여자아이 마일스톤별 평균일자>



<남자아이 마일스톤별 평균일자>



–189번

–234번

- Milestone_number : 189번
- Category : Language
- Content : 끝말 잇기를 해요
- 여자아이 평균 가능 일수 : 2062
- 남자아이 평균 가능 일수 : 2299

여자아이가 남자아이보다 237일 빠름

- Milestone_number : 234번
- Category : Recognition
- Content : 달력에서오늘날짜(월,일)를바르게가리켜요.
- 여자아이 평균 가능 일수 : 2716
- 남자아이 평균 가능 일수 : 2334

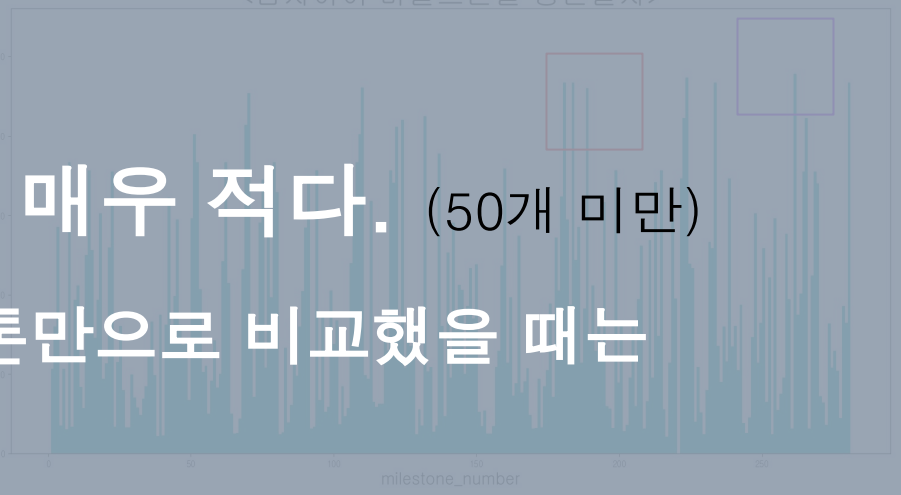
남자아이가 여자아이보다 382일 빠름

데이터 분석 성별 비교

※ 성별에 따른 마일스톤(행동) 가능일자 비교

<여자아이 마일스톤별 평균일자>

<남자아이 마일스톤별 평균일자>



해당 마일스톤의 응답 데이터 개수가 매우 적다. (50개 미만)

1000개 이상의 응답 데이터를 가지는 마일스톤만으로 비교했을 때는

성별에 따라 큰 차이가 보이지 않았다.

–234번

전체적으로 남자아이가 여자아이보다 약 11일정도 평균 가능일자가 빨랐음

- Milestone_number : 189번
  - 여자아이 평균 가능 일수 : 2062

여자아이가 남자아이보다 237일 빠름

- Milestone_number : 234번
  - 여자아이 평균 가능 일수 : 2716
  - 남자아이 평균 가능 일수 : 2334

남자아이가 여자아이보다 382일 빠름

따라서 활용할 수 있는 유의미한 결과를 도출시키기 위해서는

충분한 양의 데이터가 존재해야 한다.

알고리즘 개발 월령별 평균

코드분석 1 #함수 정의 (카플란마이어)

```
def survival_function(self, temp):

  kmf = KaplanMeierFitter()
   kmf.fit_left_censoring(
            temp['done_day_after_birth'], temp['event'])

   mid_st = kmf.median_survival_time_
   mid_avg = self.mid_time.append(mid_st)
```

생후일수(done_day_after_birth)와 event 컬럼 활용

코드분석 2 #전체 데이터 생존함수 계산

```
def survival_function_all (self):
   range_input = self.data_df['unique_number'].max()

   for i in range(range_input):
       if i == 220:
           continue

       temp = self.data_df.groupby('unique_number').get_group(i +
1).copy()
       self.start_date.append(temp['done_day_after_birth'].min())
       self.due_date.append(temp['done_day_after_birth'].max())
       self.std_date.append(temp['done_day_after_birth'].std())

       survival = self.survival_function(temp)
       self.milestone_number.append(i+1)

   return survival
```

알고리즘 개발 월령별 상위 퍼센트

ex) 마일스톤 고유 넘버 1번을 500일에 도달했을 경우

=====> 마일스톤 넘버 입력 : 1

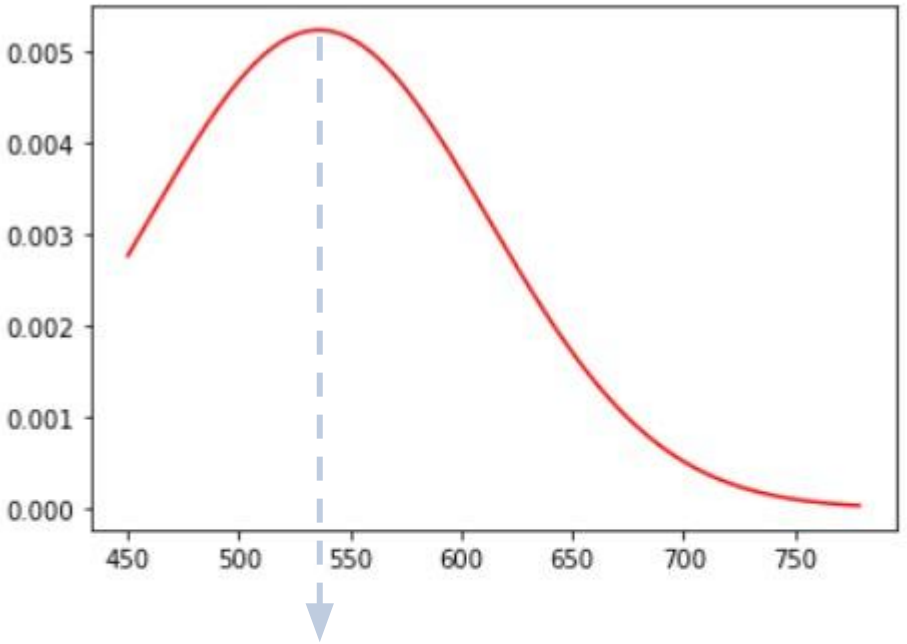
=====> 일자 입력 : 500

rv = stats.norm(536, 76.1430) # stats.norm(average, std)

카플란마이어 함수를 통해 구한 평균과
std 함수를 통해 구한 표준편차를 입력해 정규분포 생성

백분율 = rv.cdf(500) # rv.cdf(day)

확인하고 싶은 일자를 입력해 최종 백분율 도출



평균 536일

1 번의 평균 가능 일자 : 536 일

500 일은 상위 31.82% 입니다.

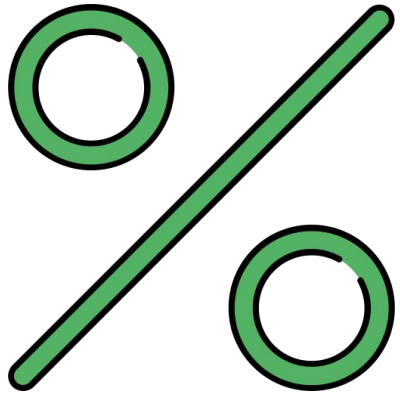
알고리즘 개발 요약

데이터 전처리

‘닥터아이‘앱에서 제공한 추출된
엑셀 데이터를 파이썬으로 전처리

각 마일스톤별 평균 구하기

생존분석을 활용하여
마일스톤별 평균 분석

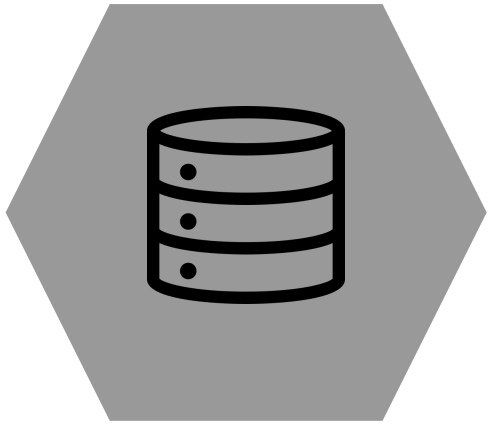


각 마일스톤별 상위 % 구하기

분석된 데이터들을 활용하여
각 user별의 상위 % 분석

알고리즘 개발 업데이트

일정량의 데이터가 들어 왔을 때 한꺼번에 업데이트하는**Batch 처리 방식** 사용



전처리 된 원래의 데이터

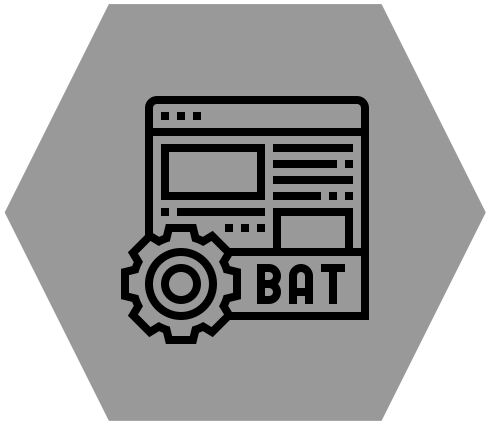
전처리 완료 후

1~8월의 데이터

새로운 데이터

전처리 완료 후
새로운 데이터 추가

9월의 데이터



합쳐진 하나의 데이터

기존 데이터와 새로운 데이터가
합쳐진 하나의 데이터

1~9월의 데이터

최종 결과 결과, 시연

ex) 168번

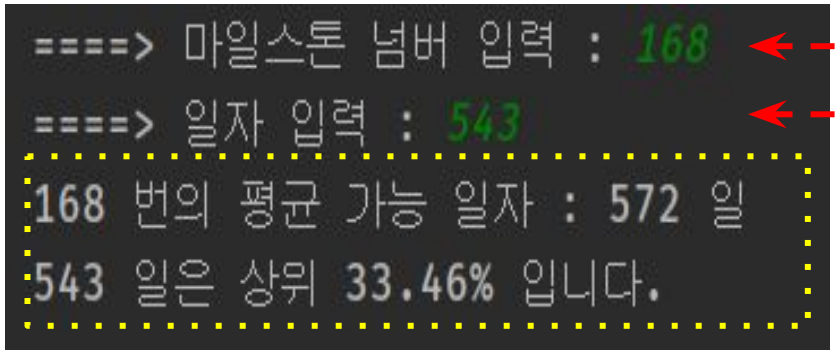
발달 퀘스트

나,"이것","저것"같은대명사를사용할수있어요 .

1 2 3 4

잘 하는 편이다.

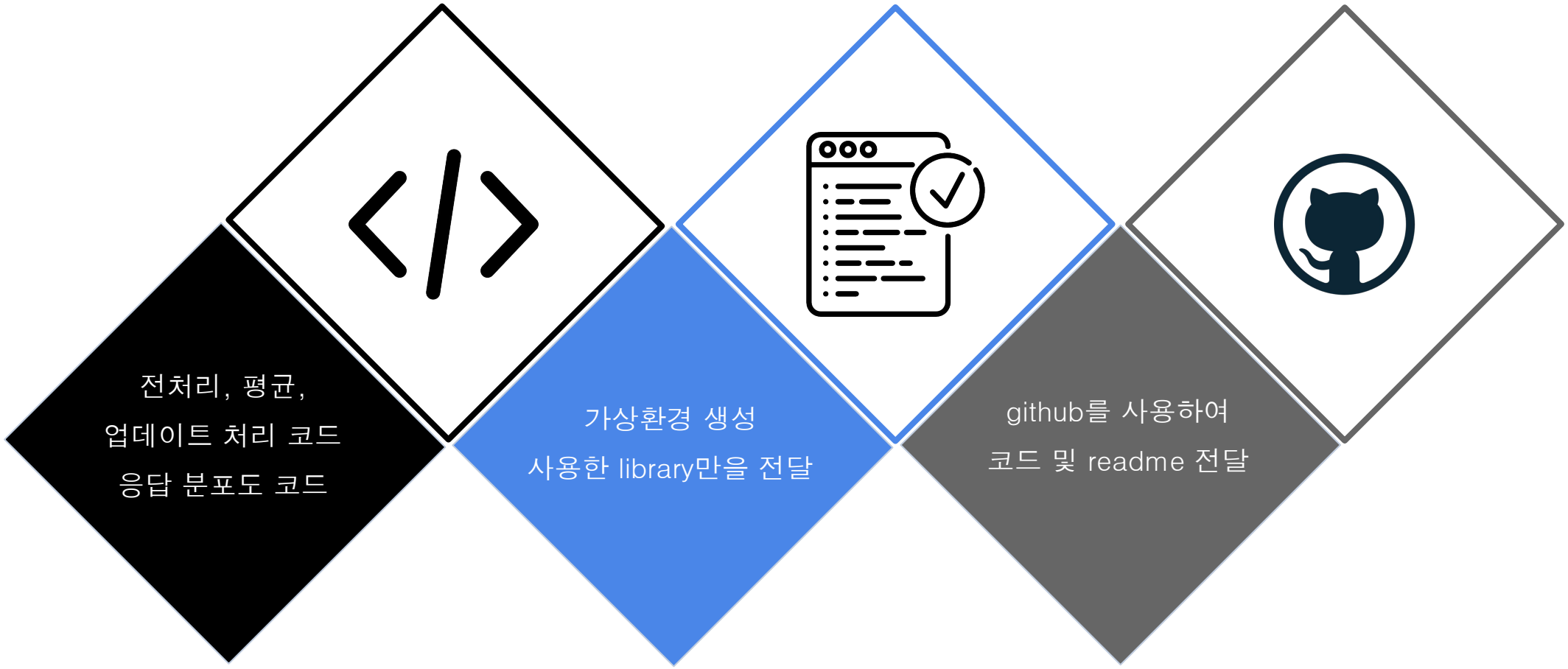
평균적으로 아이들이 걷게 되는 시점은 n일이다 or n일에 걸으면 상위 n%이다.



해당 마일스톤 넘버

사용자의 생후일자

최종 결과 전달자료


전처리, 평균,
업데이트 처리 코드
응답 분포도 코드
가상환경 생성
사용한 library만을 전달
github를 사용하여
코드 및 readme 전달

CONTENTS

1

데이터 분석 연구 배경

‘닥터아이’ 서비스

목표

2

데이터 분석 진행 과정

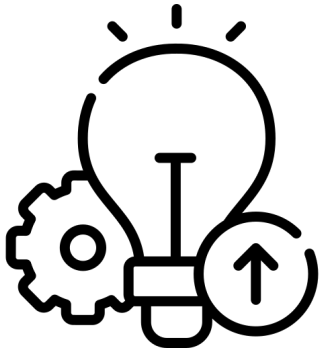
데이터 전처리

데이터 분석

알고리즘 개발

최종결과

3

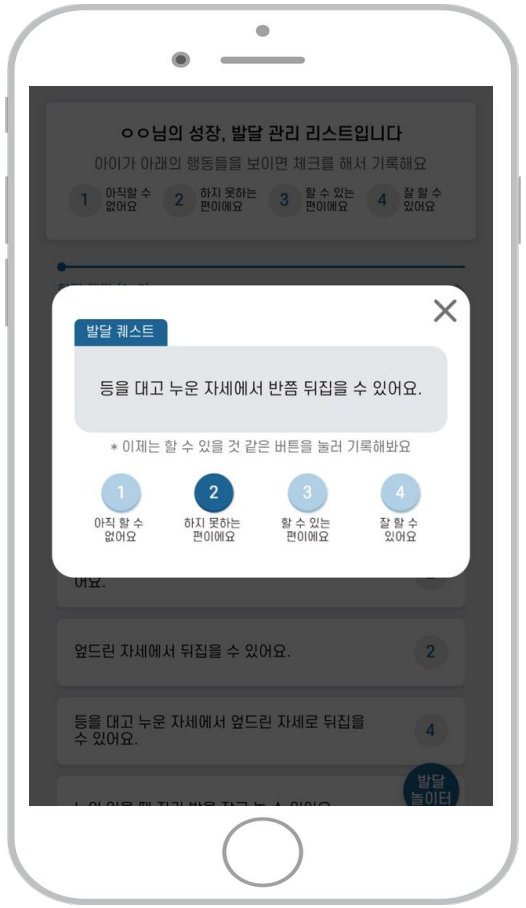


적용과 기대효과

기존 서비스와의 차별성

활용방안과 기대효과

기존 서비스와의 차별성



기존 서비스

생후일수에 해당하는 질문 제공
이를 통한 사용자의 응답

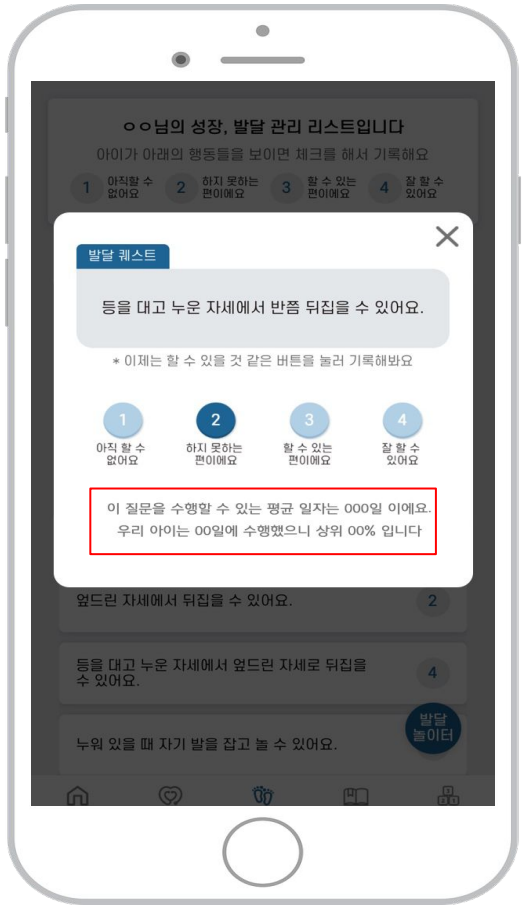
기존 서비스

해당 질문을 통해 아이가 잘 자라고 있는 지 자가 체크 가능

아쉬운 점

단, 본인의 아이 정보만 확인 가능
또래 아이와의 비교 불가능

기존 서비스와의 차별성



기대하는 것

사용자 아이의 성장 확인만이 아닌
다른아이들과의 성장 비교

수정되는 서비스

질문에 체크를 하면 하단에 응답한
아이들의 평균일자와 상위 %를 띄움

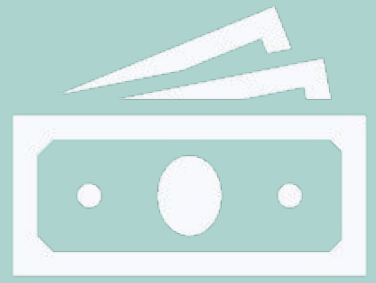
차별점

우리 아이의 성장 속도를
객관적으로 판단 할 수 있음

활용방안과 기대효과

Detailed Feedback

발달상황에 대한 자세한 피드백 제공
마일스톤 별 참고 수치 제공
(평균 가능 일자, 백분율)

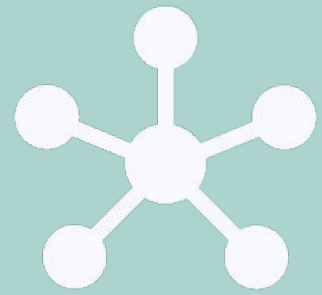


Cost Reduction

보완된 영유아 발달 검사로
외부 발달 검사 대체 효과

Extract Insight

추후 더 많은 양의 데이터가 축적 되면
더 신뢰성 있는 인사이트 도출 가능



Service Diversification

발달 상황에 따른 부가서비스 다각화

Thank you

Q&A