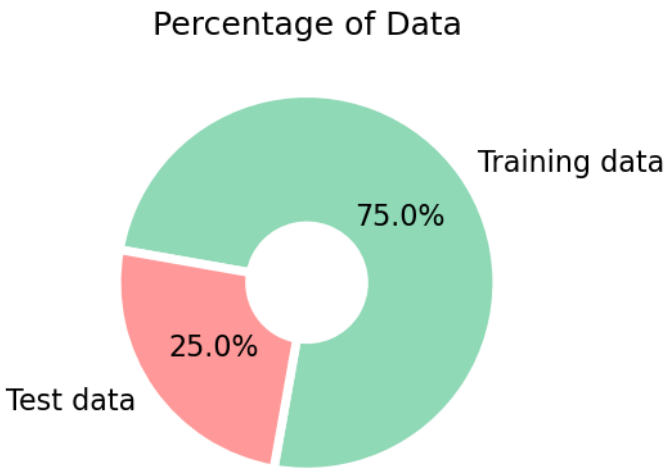
# EDA Report

## Reference

[UCI archive] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

[Kaggle] https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

## Dataset




The number of training data : 161297

The number of test data : 53766

## Dataset info

| | |
|---|---|
| Number of features | 7 |
| Total missing (%) | 0.5579% |

## Feature types

| | |
|---|---|
| Numeric | 2 |
| Categorical | 4 |
| Date | 1 |
| Dropped | 0 |

## Data analysis

전체 데이터 구성은 uniqueID를 가진 환자가 가지고 있는 증상에 필요한 약을 구입한 뒤에 특정 날짜에 review와 rating을 남김. 그리고 다른 사람이 해당 리뷰를 보고 도움이 되었는지에 대해 usefulCount feature에 점수(1점 추가)를 줌.

**Feature Info**

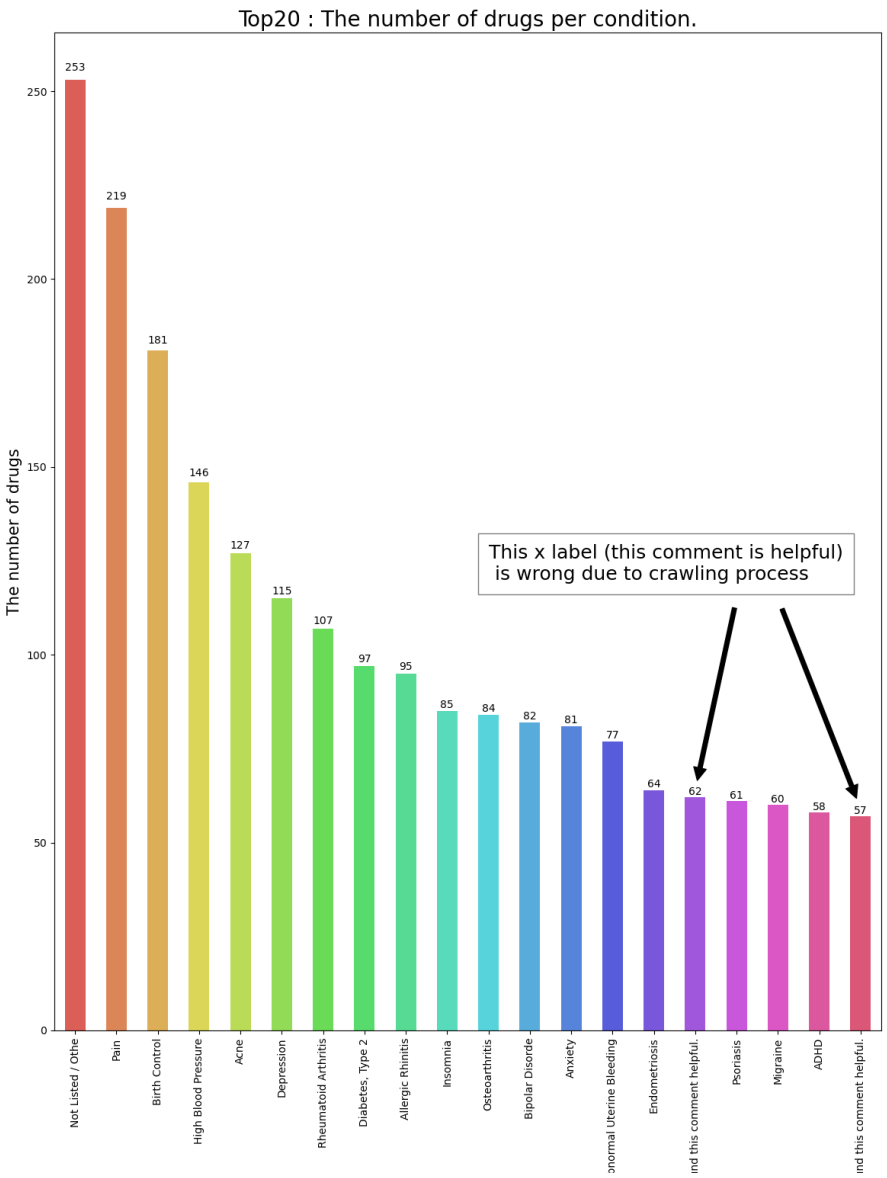
| Feature | Type | Description |
|---|---|---|
| uniqueID | Numeric | Identify individual data |
| drugName | Categorical | Name of drug |
| condition | Categorical | Name of condition |
| review | Categorical | Patient review |
| rating | Categorical | 10star patient rating |
| date | Date | Date of review entry |
| usefulCount | Numeric | Number of users who found review useful |

## Condition, drugName

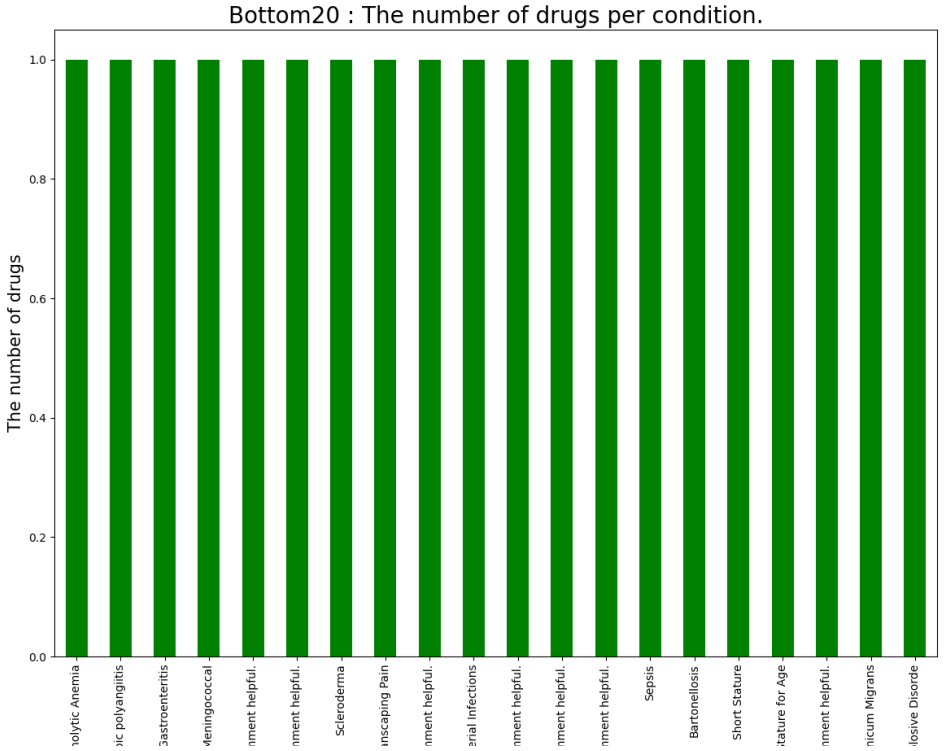
The number of unique condition : 937

The number of unique drugName : 3671

Below 4 charts show the pattern between drugName and condition

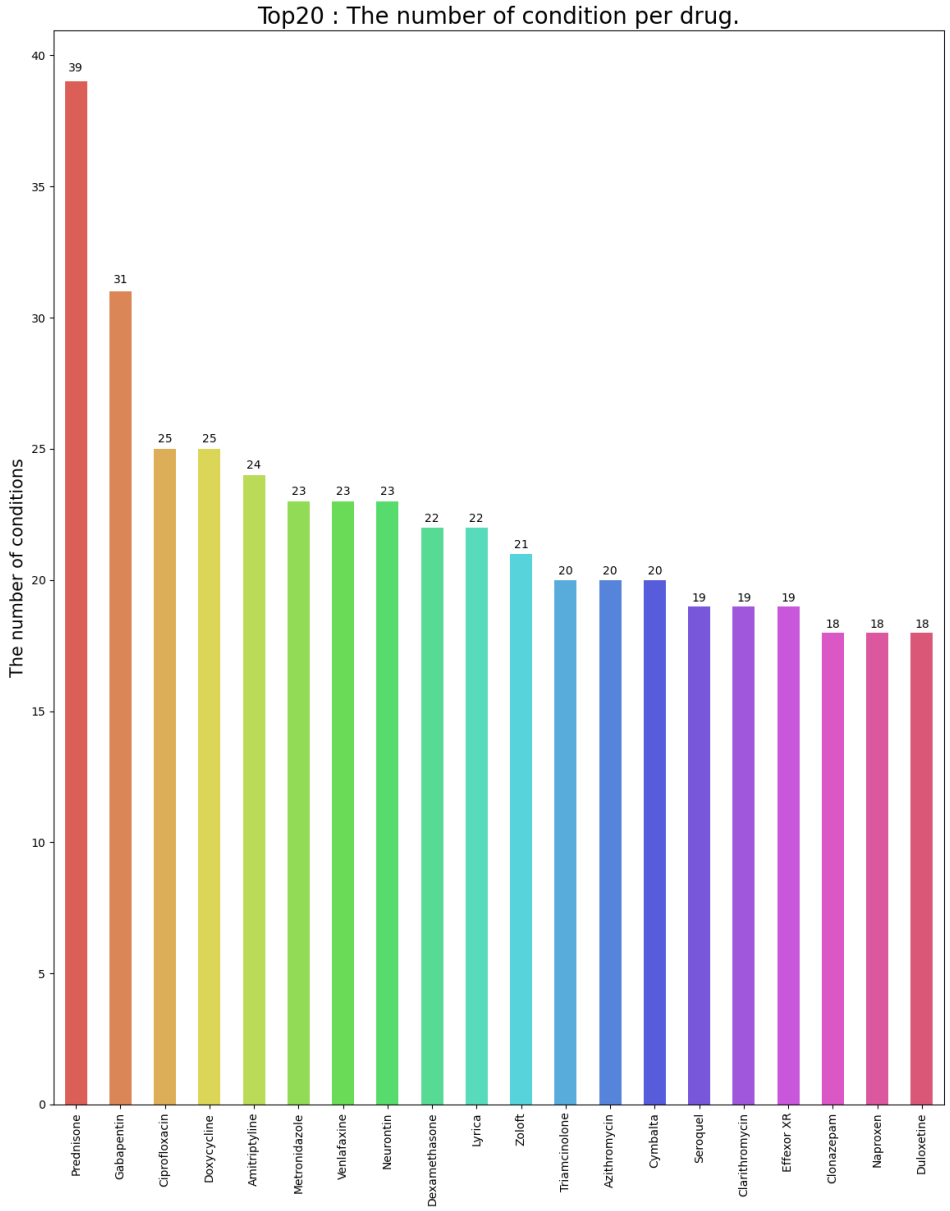



X label 중에서 html 태그로 인해 데이터가 잘못 들어가 있는 상태를 화살표로 표시함.



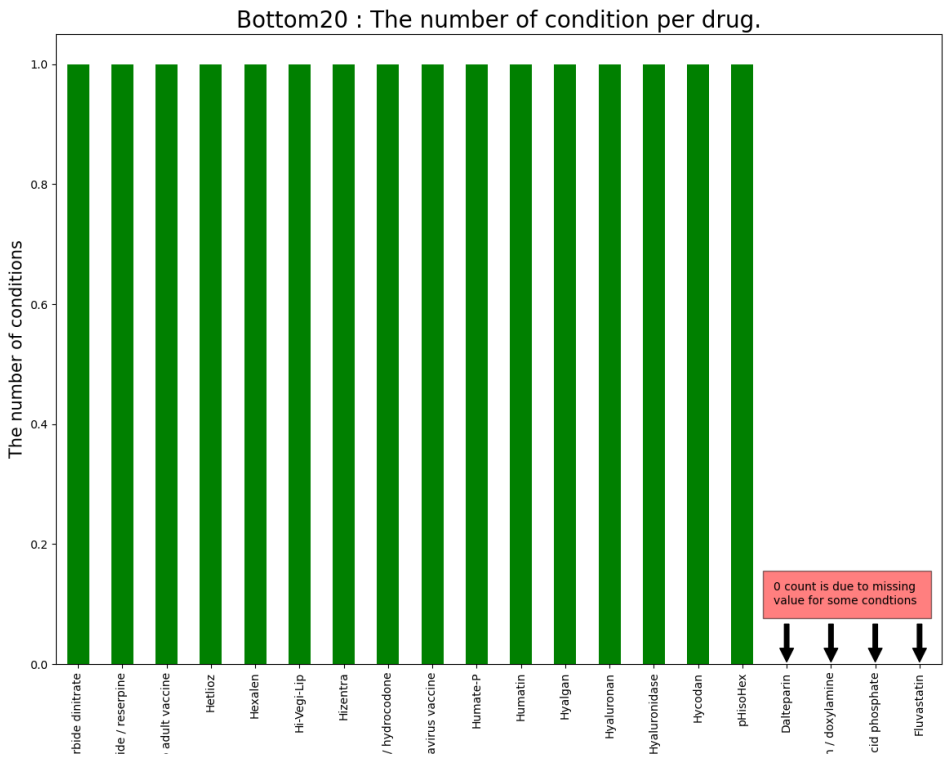

위의 그래프와 비교하여 보면, 컨디션당 사용된 약물이 250여개가 되기도 하지만, 1개씩만 사용된 약의 경우도 상당함을 확인할 수 있음.

이와 같은 경우 데이터에 편차가 발생할 수 있을 것으로 예상됨.



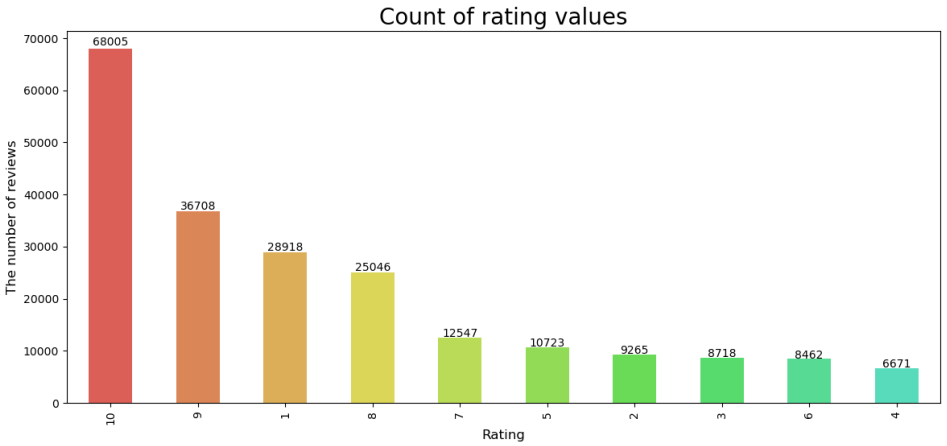

drugName과 condition feature간의 패턴을 파악하기 위해, 이번에는 반대로 drug당 증상의 개수를 분석하였음.

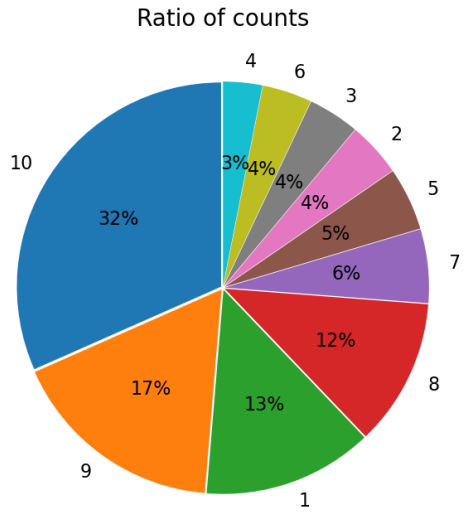
하나의 약물이 여러 컨디션에 활용되는 경우를 확인할 수 있음.




결측치가 발생한 경우도 확인할 수 있음.(약은 있는데, 컨디션이 NaN)

## rating



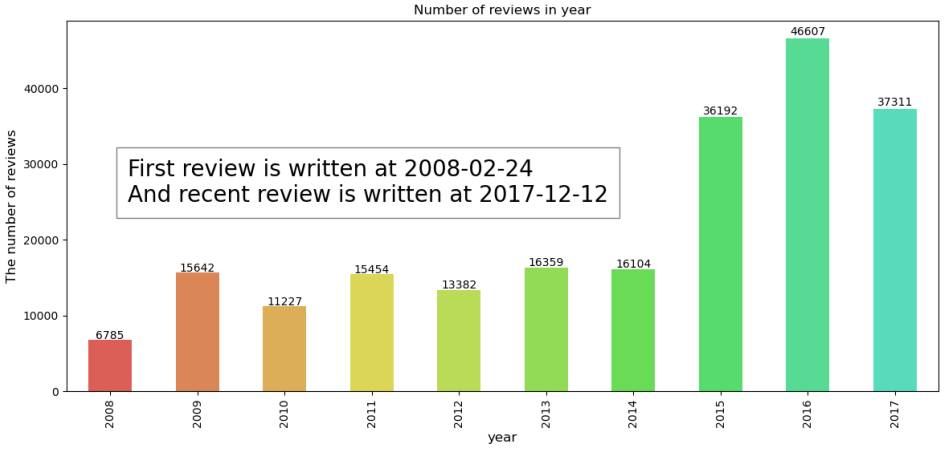



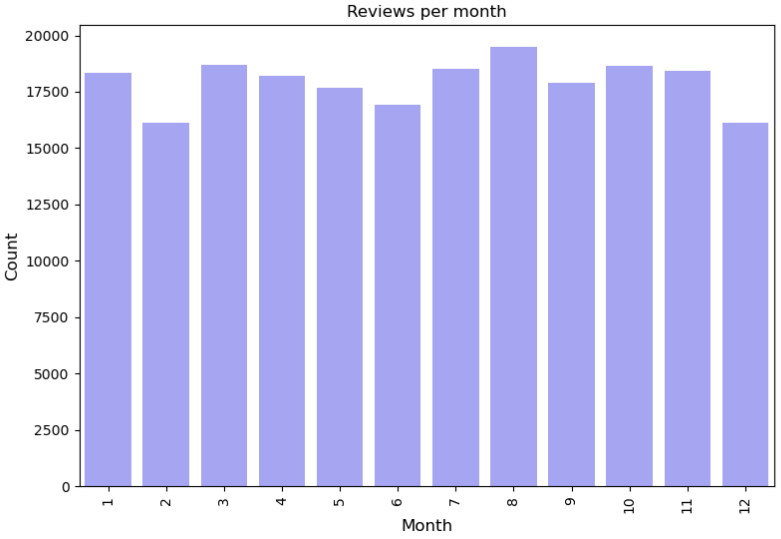

rating별 리뷰의 개수를 나타낸 도표.

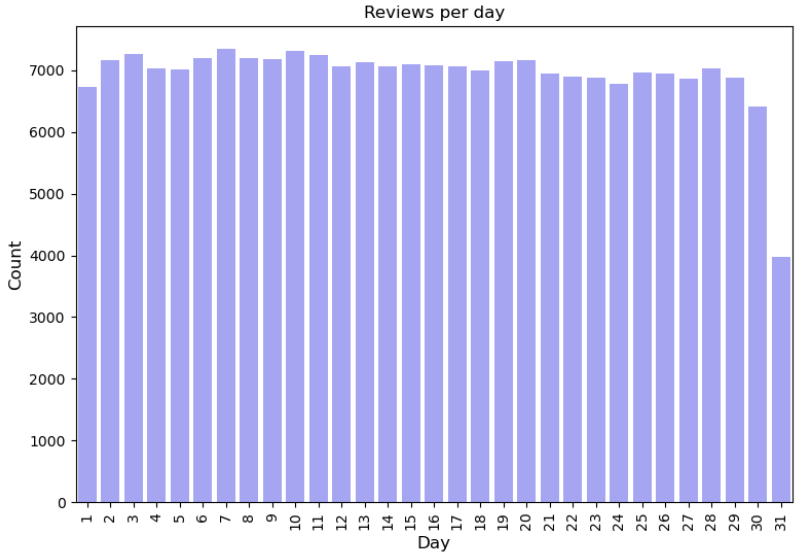
10점, 9점, 1점 순으로 리뷰가 많음을 확인할 수 있음.

점수가 극단적인 경우의 리뷰가 많고, 중간 점수의 리뷰는 적음.

## date


Number of reviews in year
First review is written at 2008-02-24
And recent review is written at 2017-12-12
The number of reviews
year
6785
15642
11227
15454
13382
16359
16104
36192
46607
37311
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017


Reviews per month
Count
Month

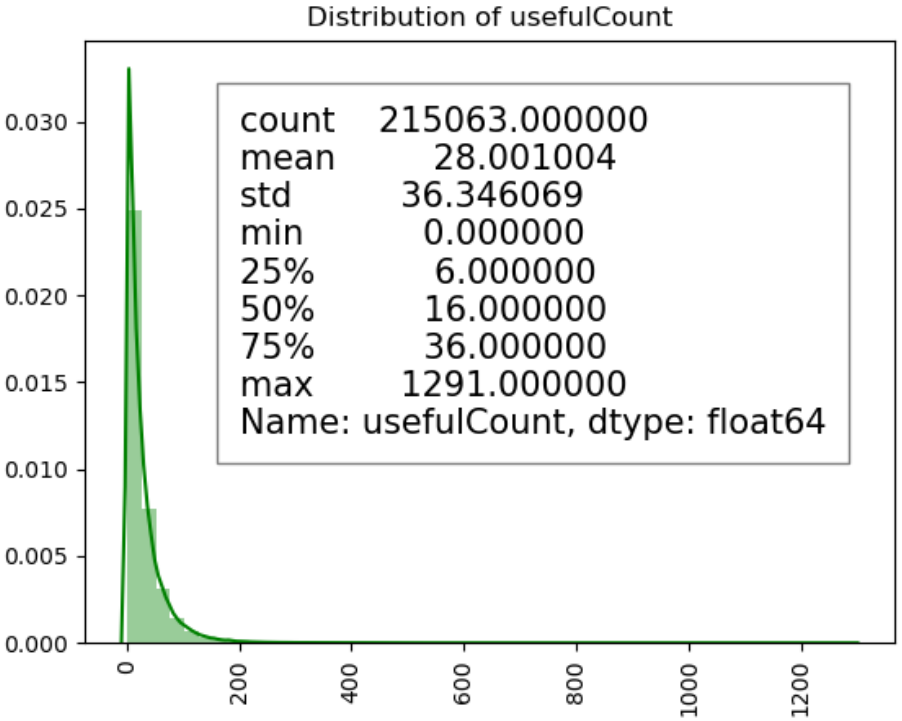



해가 갈수록 작성되는 리뷰의 수가 증가하는 모습

리뷰 작성일이나 날짜에 따른 편향은 거의 존재하지 않는 것으로 보임

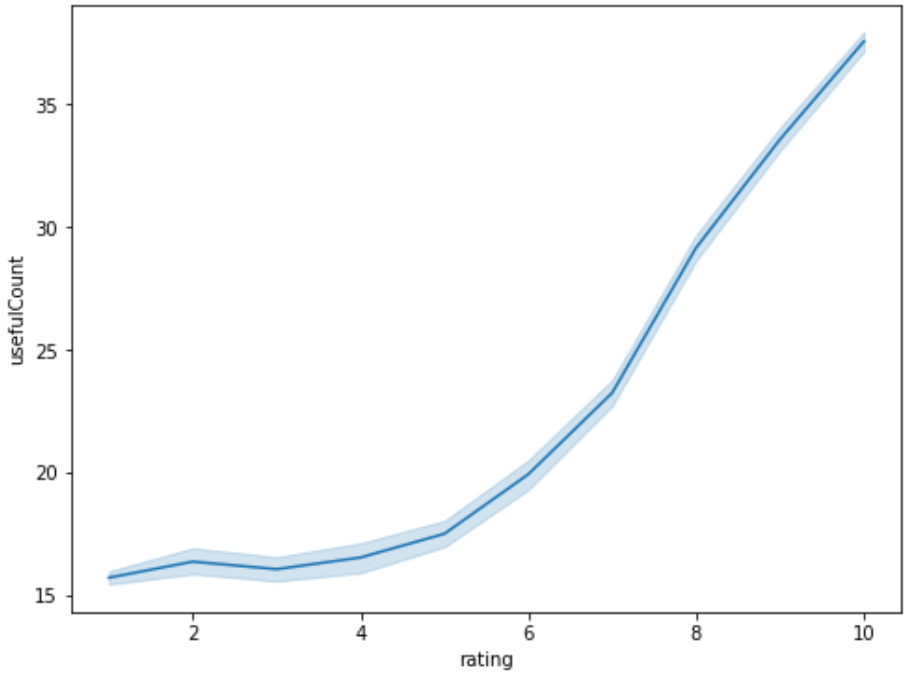
단, 31일이 있는 달은 적기 때문에, 31일에 쓰여진 리뷰는 다른 날(day)보다 리뷰의 개수가 적음

## usefulCount




usefulCount는 분포 범위가 매우 넓으며 분포 편차도 매우 큼

Relationship between usefulCount and rating




대체로 rating이 높은 리뷰가 usefulCount가 높은 경향이 발견됨

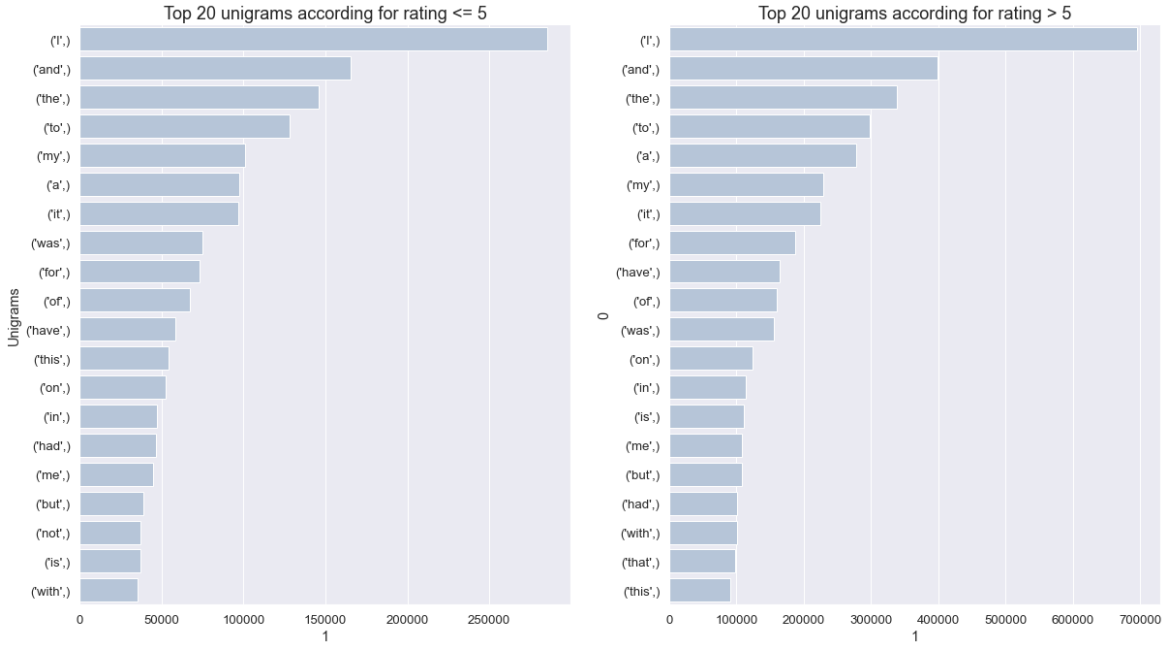
사람들이 직접 먹을 약을 찾기 위해 부정적인 리뷰보다 긍정적인 리뷰에 더 호응한것으로 추정 (positive screening)

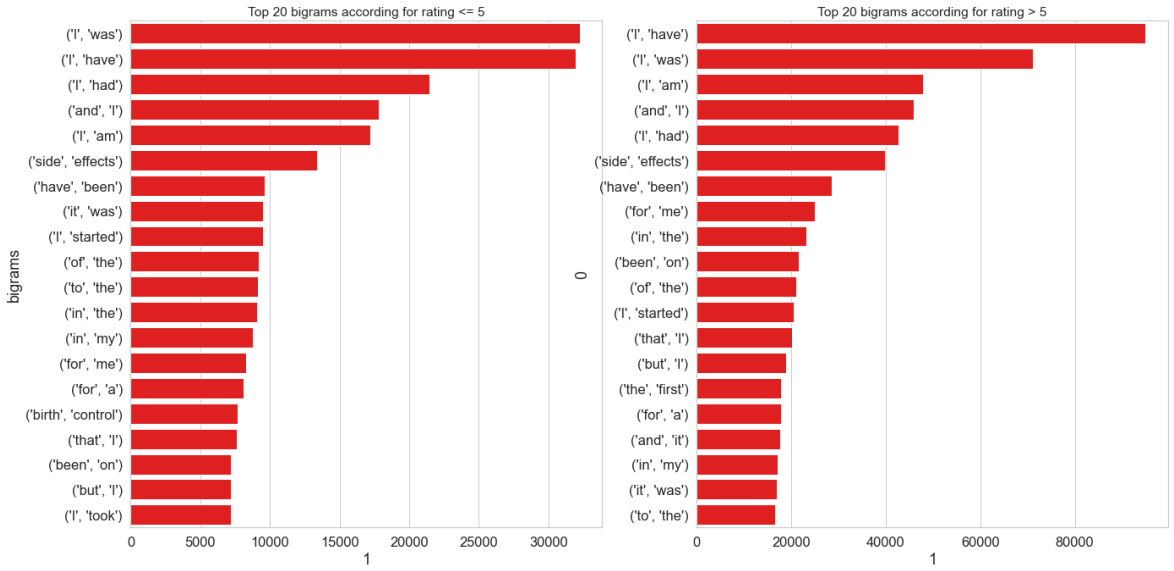
## review

리뷰 데이터에는 html태그와, 에러 문자, 대문자로 강조된 단어, 괄호 안에 감정표현이 포함되어 있음.

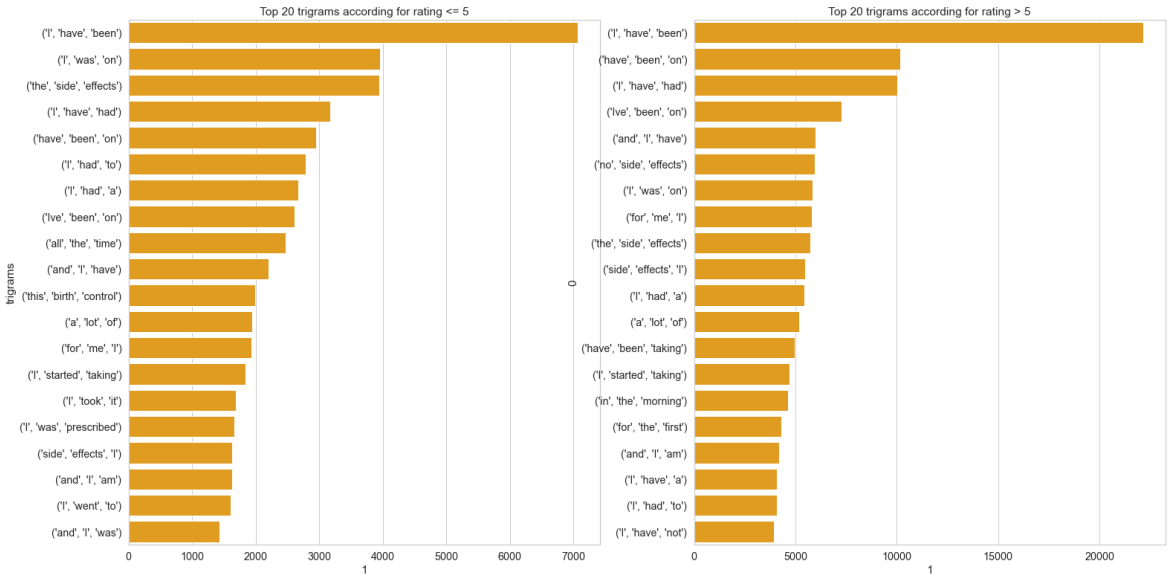
"If not for this antibiotic my husband would have been dead several times over. Incidentally **I'm** a Registered nurse. I see where a number have people have attributed wild **"**side effects**'** to this drug. Many are confusing symptoms of their illness with side effects of the antibiotic. Just because you experience something unusual for you when your sick doesn**'**t mean its the fault of the antibiotic. Being ill is an abnormal state for ones body yet some of these people seem to think that any effect of the illness they weren**'**t expecting **MUST BE** attributable to something else and often they blame the treatment rather than the illness. If you**'**ve been prescribed Levaquin I urge you to take many of these **'**reviews**'** with more than a grain of salt! An **EXCELLENT** drug!"

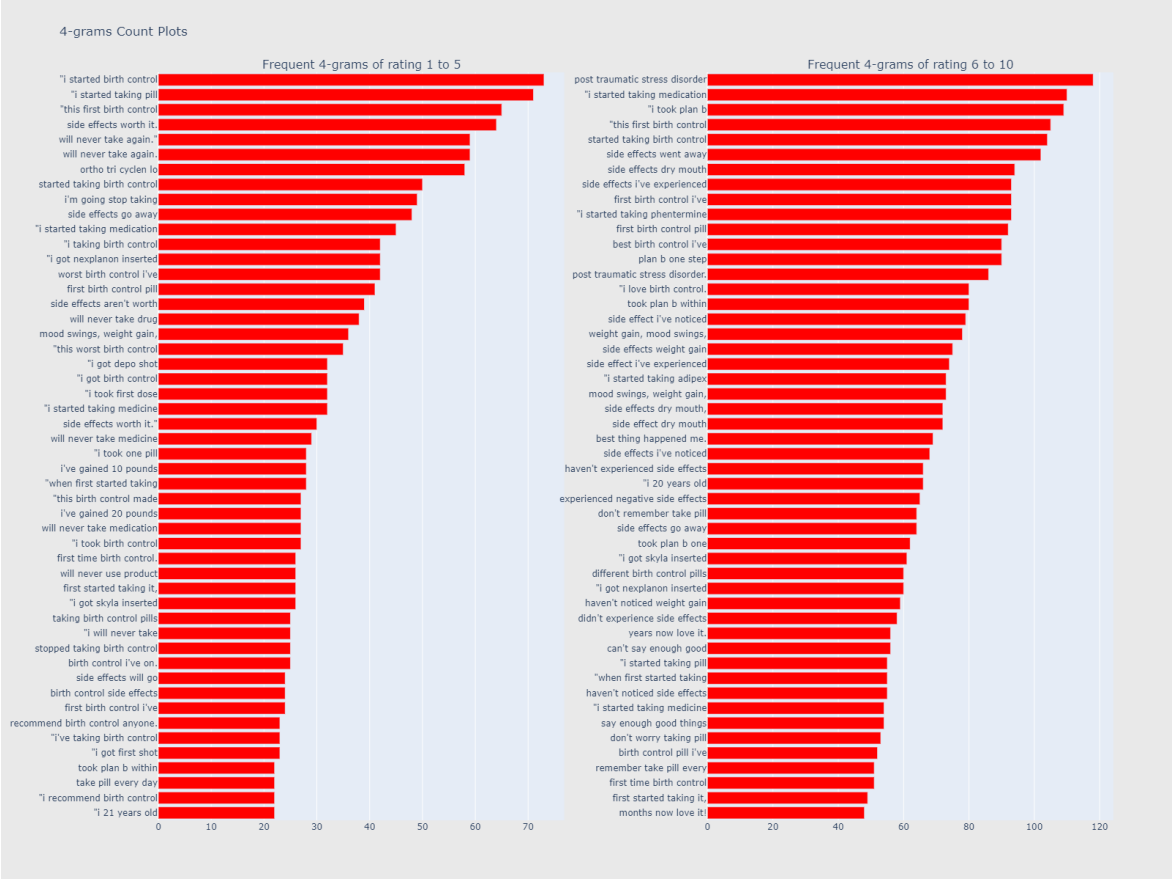
rating을 기준으로 5점 이하는 negative, 그 외는 positive로 분류함
이후 단어들을 1-gram, 2-gram, 3-gram, 4-gram으로 나누어 빈도수를 분석함






1, and 2- gram shows same distribution both positive(rating>5) and negative(rating<=5) at top 5






1, 2, 3-gram에서는 불용어가 많이 포함되어있어 유의미한 차이를 확인할 수 없었으나, 4-gram에서는 유의미한 차이가 나타남.

## 데이터셋의 한계점 및 생각한 대안

저희 5명이서 각자 나뉘어 kaggle에 1위 부터 5위 까지 랭크된 포스트를 분석 하였습니다. 그러나 모든 모델이 비슷하게 리뷰를 이용해서 감성(positive, negative)분석을 진행하였고 최종 예측으로 rating을 맞추는 모델들이었습니다.

이미 진행된 모델들이 모두 리뷰를 사용하여 평점을 예측하는데 목적이 있어서, 리뷰데이터가 없다면 평점을 예측하기 어렵다는 단점이 있습니다. 이는 챗봇으로 사람들에게 증상을 입력받아 약을 추천해주는 서비스에는 적합하지 않은것 같습니다.

대안

1. 리뷰데이터로 단어를 학습 ex) (input) 배가 아프다 -> 학습된 데이터에서 유사도가 높은 약을 추천. (Word2Vec 활용)
2. 다른 추가적인 데이터셋 찾기
3. 챗봇에만 집중. rating+머신러닝을 활용한 점수 개발로 순위 매기기
   -> 미리 증상별로 준비된 순위에 대해 단순 쿼리로 결과 보여주기
4. 웹사이트에서 블로그, 카페에서 의약품에 대한 후기 크롤링
   -> RNN 모델에 적용하여 리뷰만으로 평점 예측하여 데이터 셋 구축