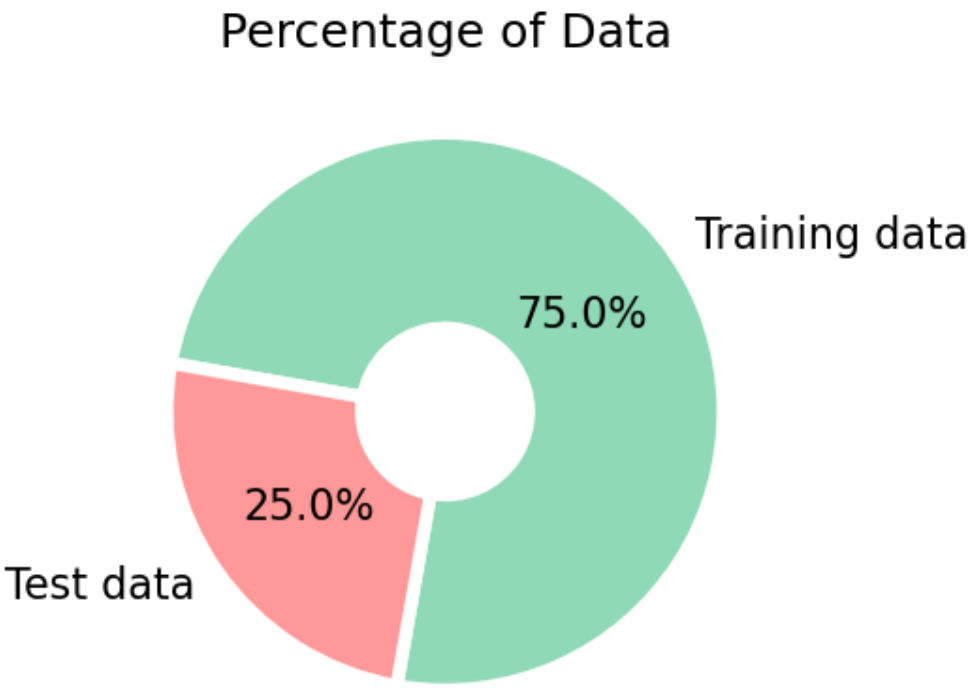
EDA Report

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		Feature types	
Number of features	7	Numeric	2
Total missing (%)	0.5579%	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

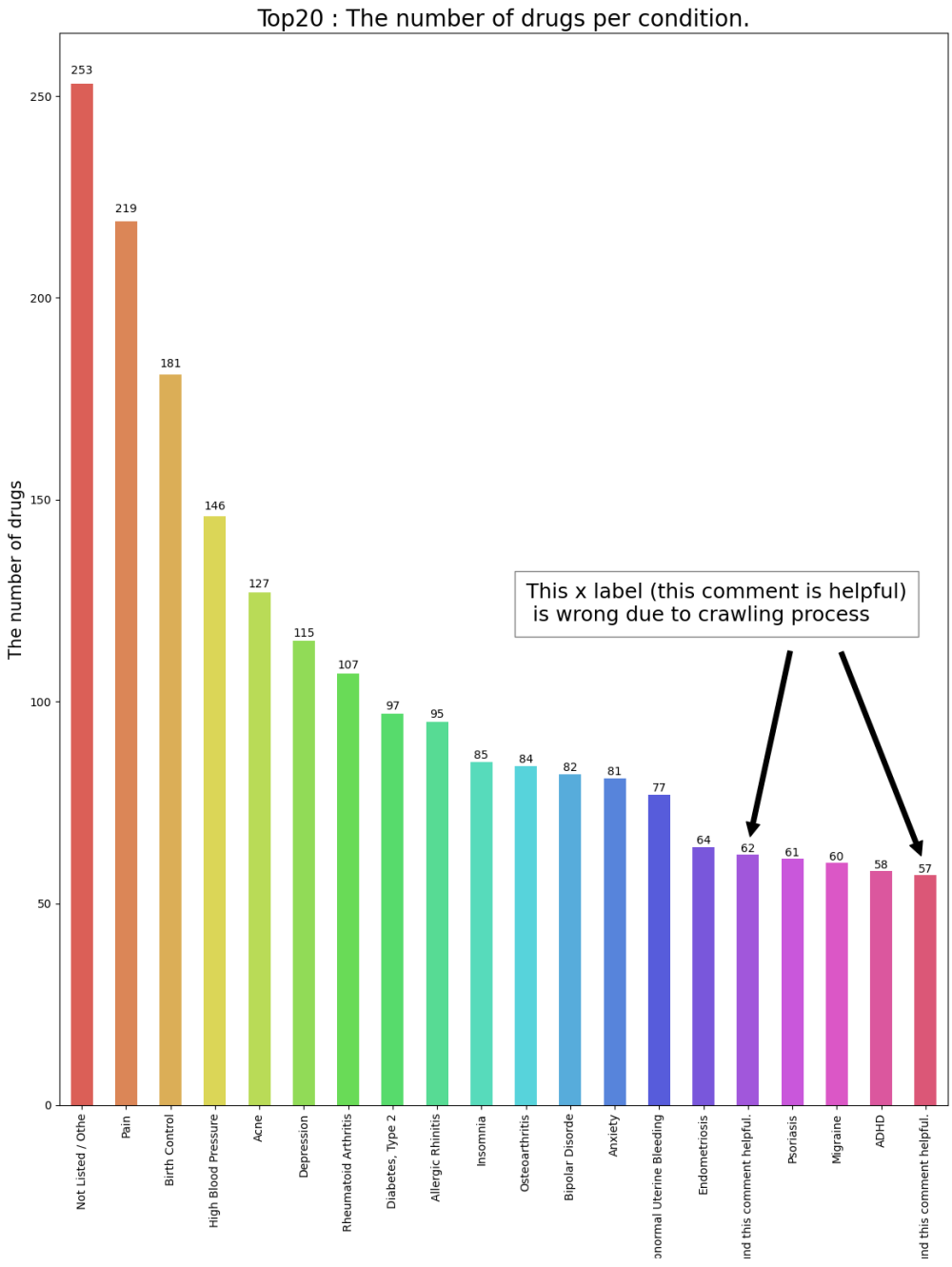
Feature	Type	Description
uniqueID	Numeric	Identify individual data
drugName	Categorical	Name of drug
condition	Categorical	Name of condition
review		Patient review
rating	Categorical	10star patient rating
date	Date	Date of review entry
usefulCount	Numeric	Number of users who found review useful

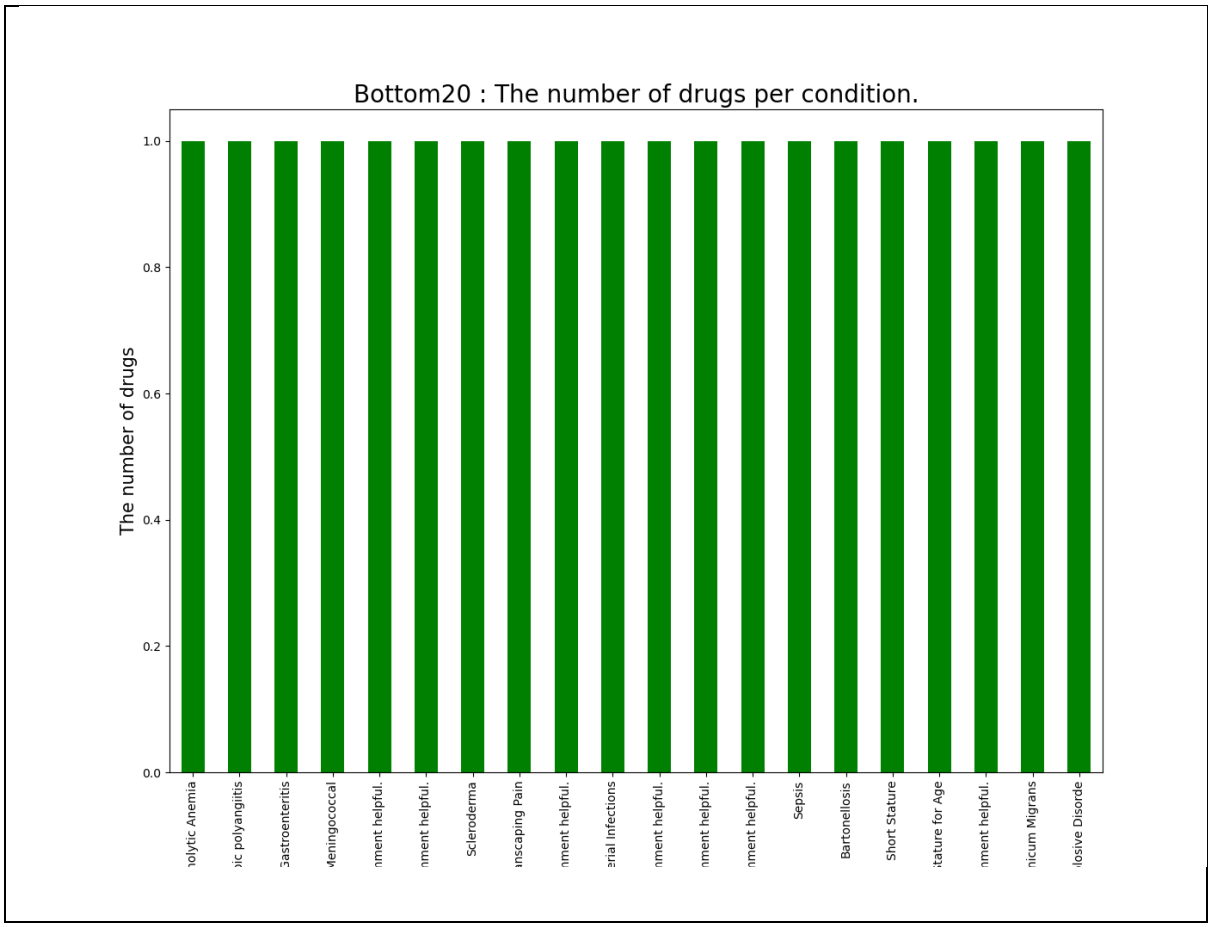
Condition, drugName

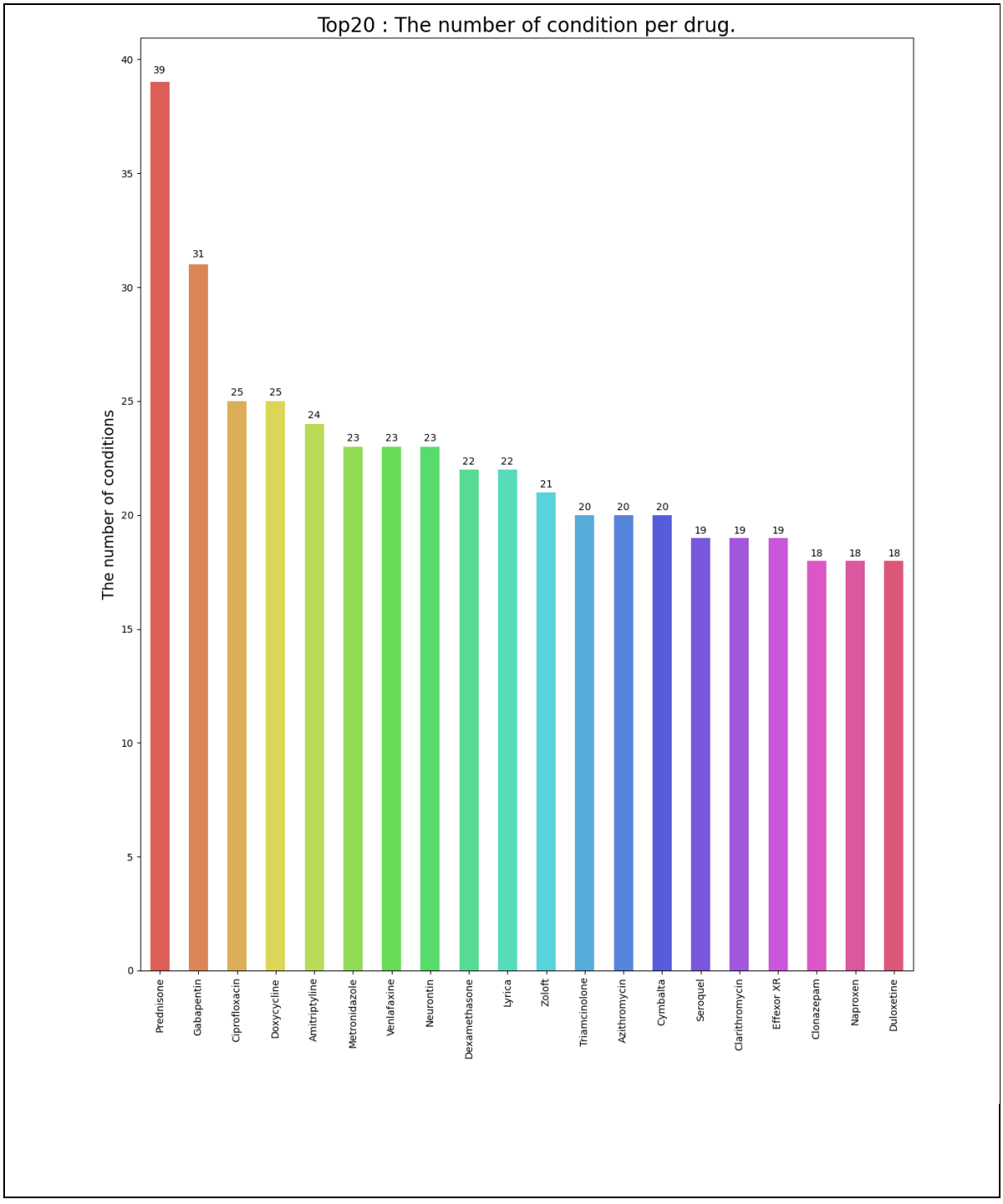
The number of unique condition **:** 937

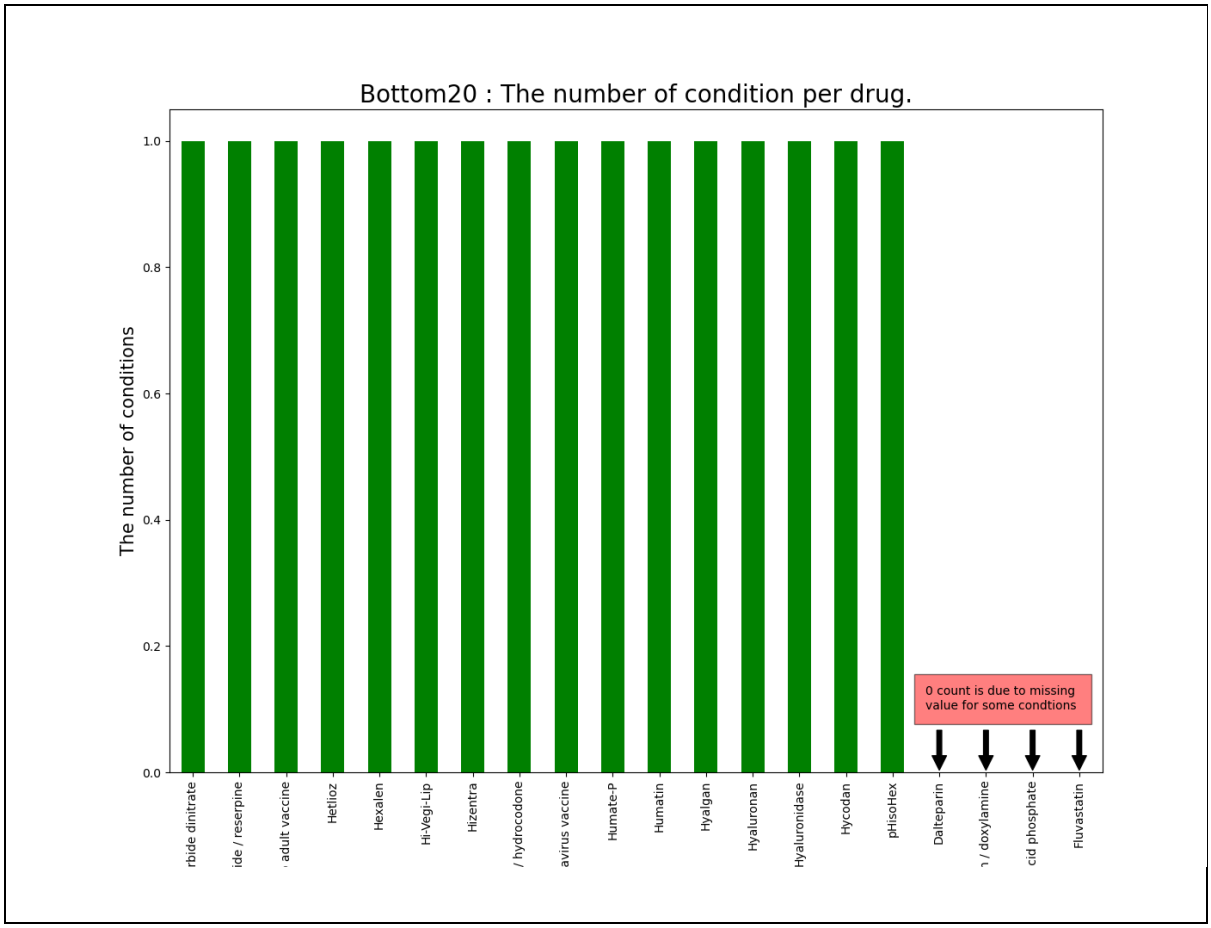
The number of unique drugName **:** 3671

Below 4 charts show the pattern between drugName and condition

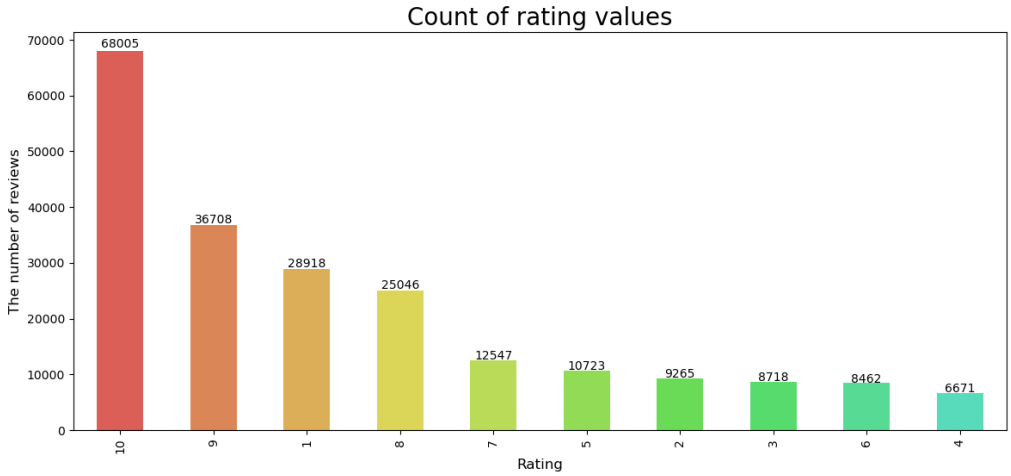


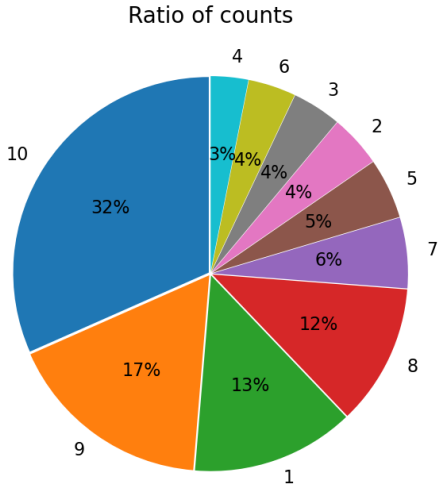

Bottom20 : The number of drugs per condition.
The number of drugs
1.0
0.8
0.6
0.4
0.2
0.0
Sclerodema
Sepsis
Bartonellosis
Short Stature

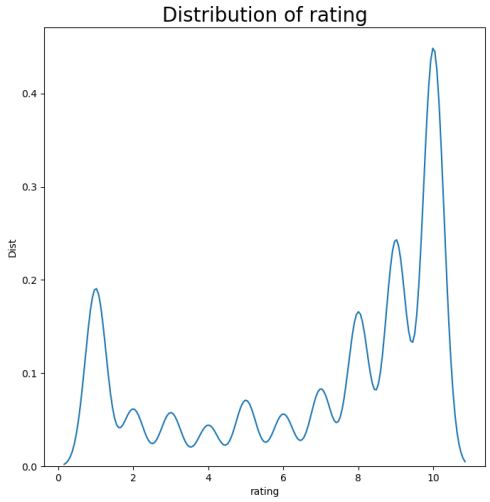

Top20 : The number of condition per drug.
The number of conditions
0
5
10
15
20
25
30
35
40
39
31
25
25
24
23
23
23
22
22
21
20
20
20
19
19
19
18
18
18
Prednisone
Gabapentin
Ciprofloxacin
Doxycycline
Amitriptyline
Metronidazole
Venlafaxine
Neurontin
Dexamethasone
Lyrica
Zoloft
Triamcinolone
Azithromycin
Cymbalta
Seroquel
Clarithromycin
Effexor XR
Clonazepam
Naproxen
Duloxetine

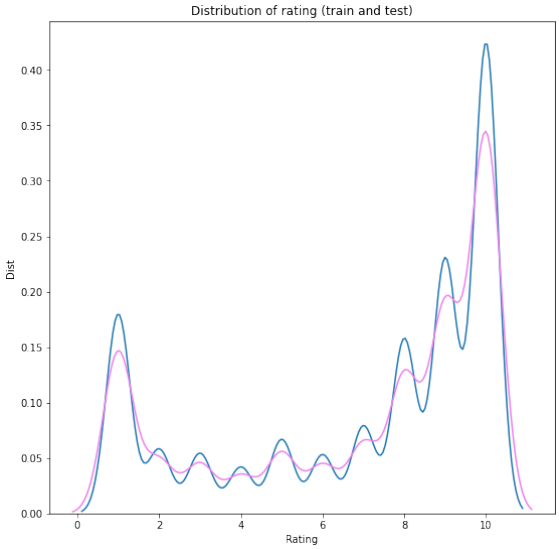

Bottom20 : The number of condition per drug.
The number of conditions
1.0
0.8
0.6
0.4
0.2
0.0
0 count is due to missing value for some condtions
Hetlioz
Hexalen
Hi-Vegi-Lip
Hizentra
Humate-P
Humatin
Hyalgan
Hyaluronan
Hycodan
pHisoHex
Dalteparin
Fluvastatin

rating

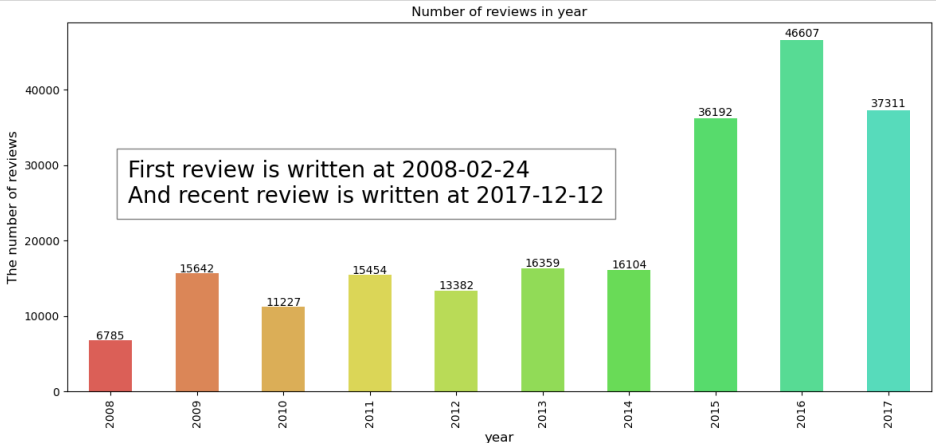

Count of rating values
The number of reviews
Rating
68005
36708
28918
25046
12547
10723
9265
8718
8462
6671
10
9
1
8
7
5
2
3
6
4

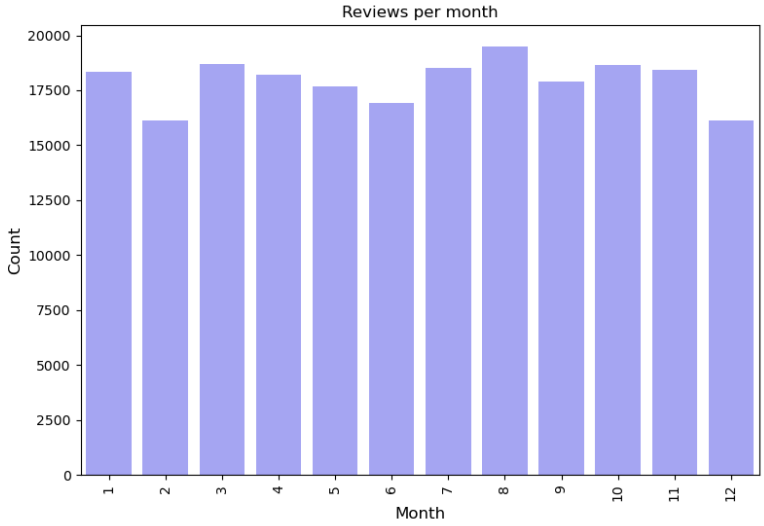

Ratio of counts
10
4
6
3
2
5
7
8
9
1
32%
3%
4%
4%
4%
5%
6%
12%
13%
17%

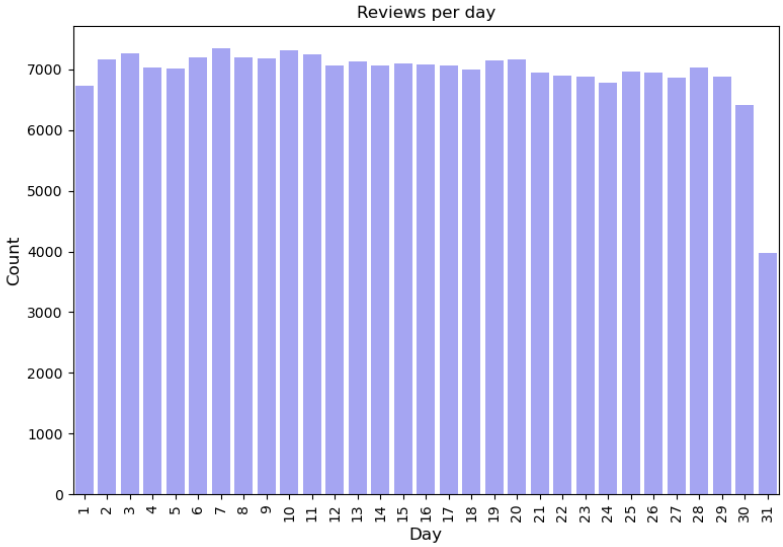

Distribution of rating
Dist
rating


Distribution of rating (train and test)
Dist
Rating

date

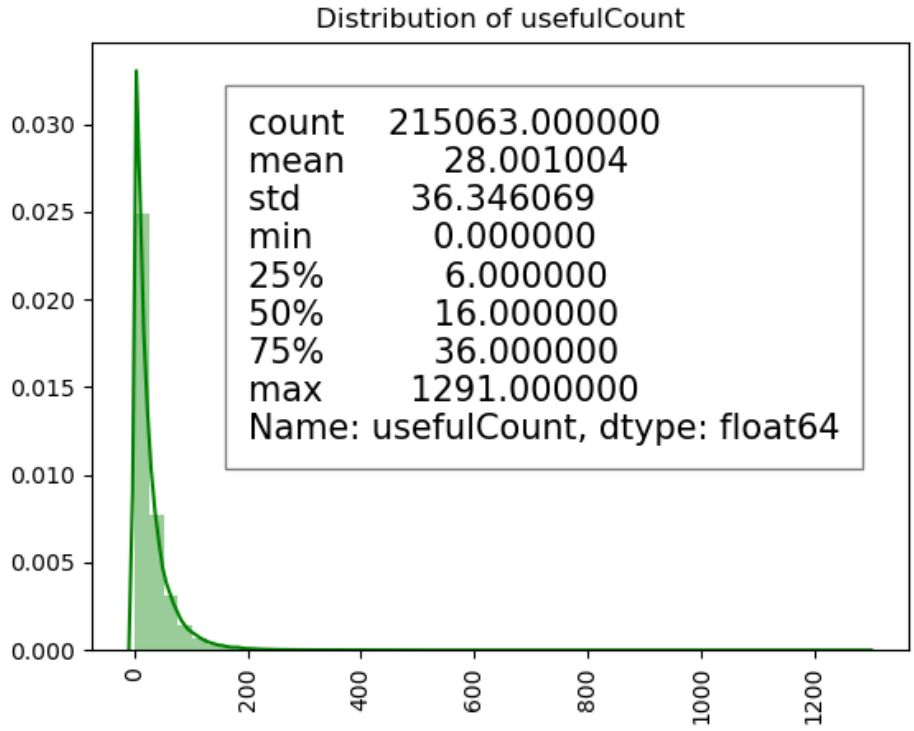






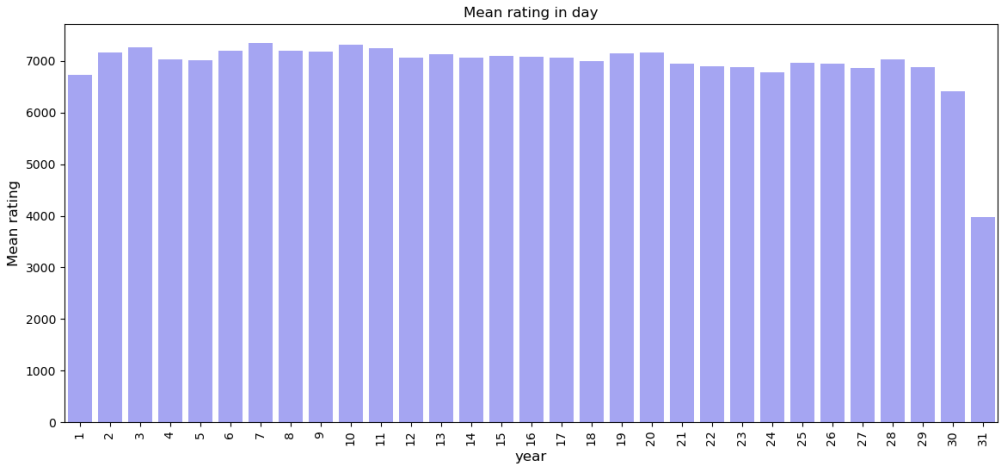
->31일이 있는 달은 적기 때문에, 31일에 쓰여진 리뷰는 다른 날(day)보다 리뷰의 개수가 적음

usefulCount

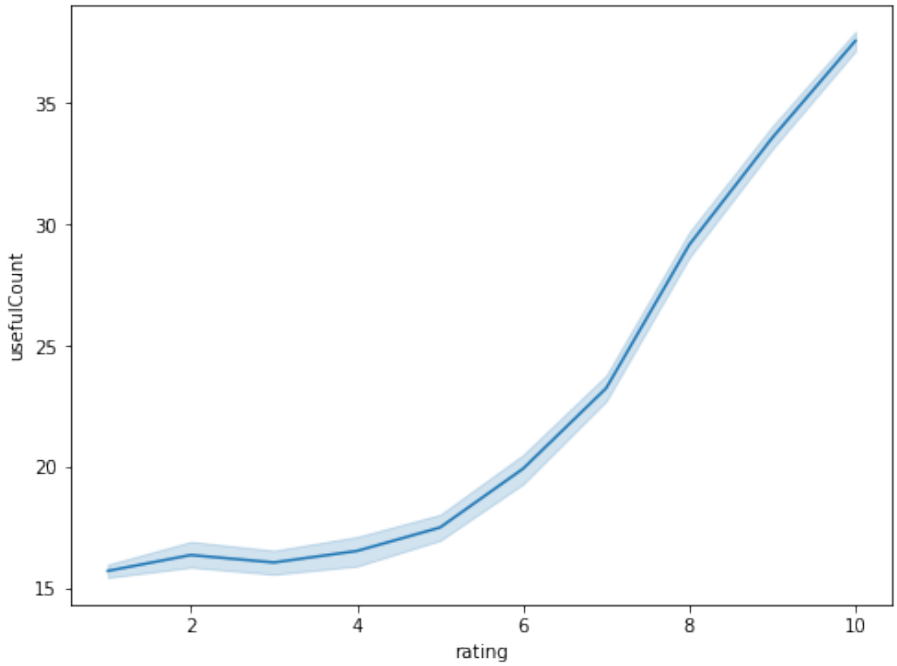


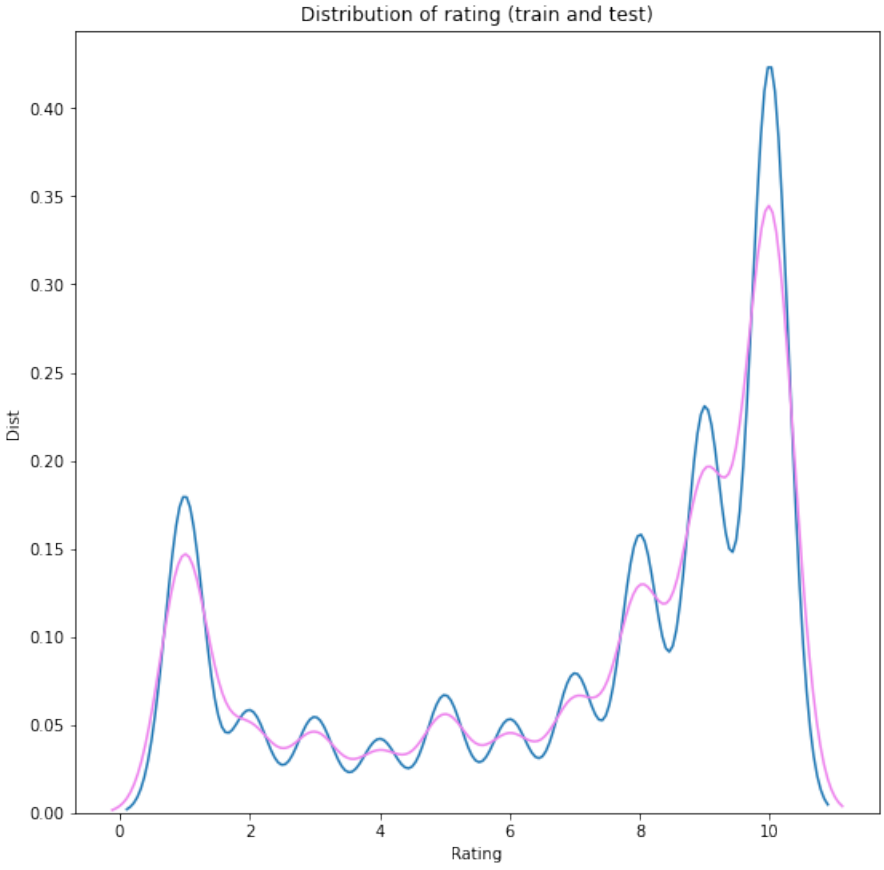
The usefulCount shows extreme between min and max. Also, standard deviation is high.

Below chart shows the relationship between usefulCount and date



Relationship between usefulCount and rating





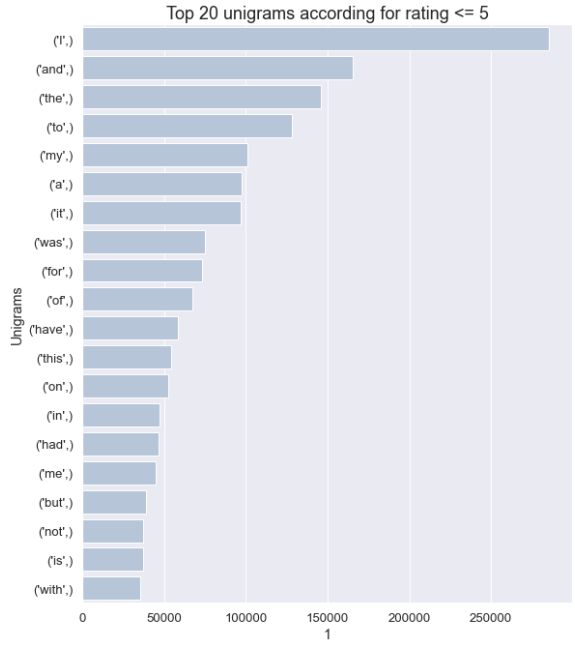
->이거 한페이지에 따로 각 feature별로 분포도 다 같이 나타내면 될 듯

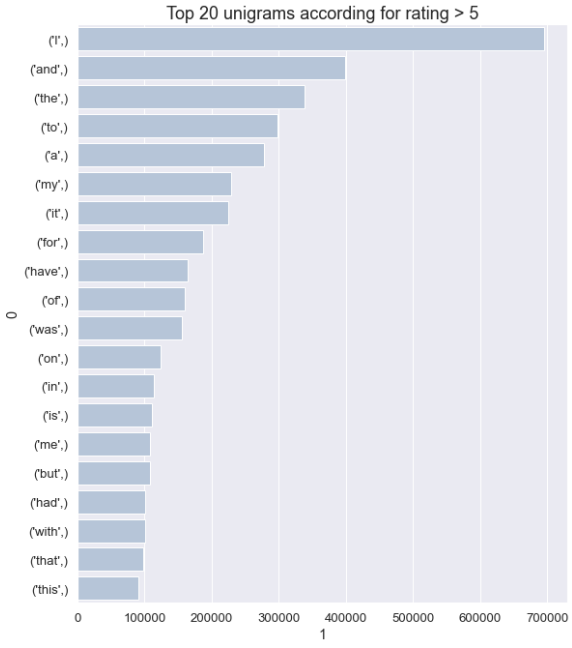
review

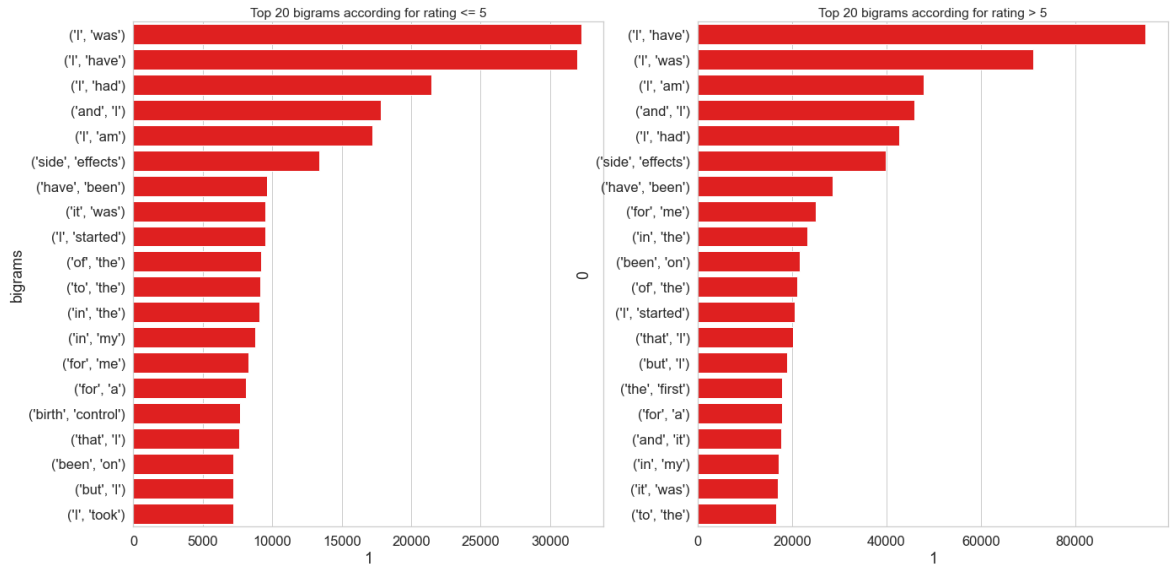
There is some html tag, error strings, uppercase word, and emotional parenthesis in reviews

"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!"

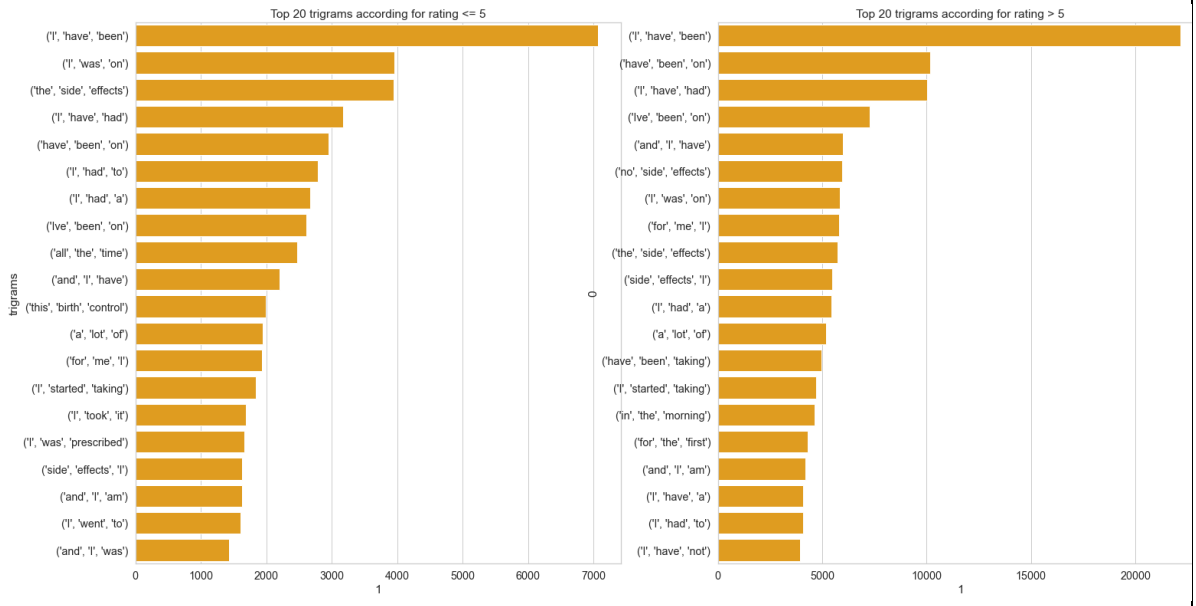
Below 3 charts show N-gram

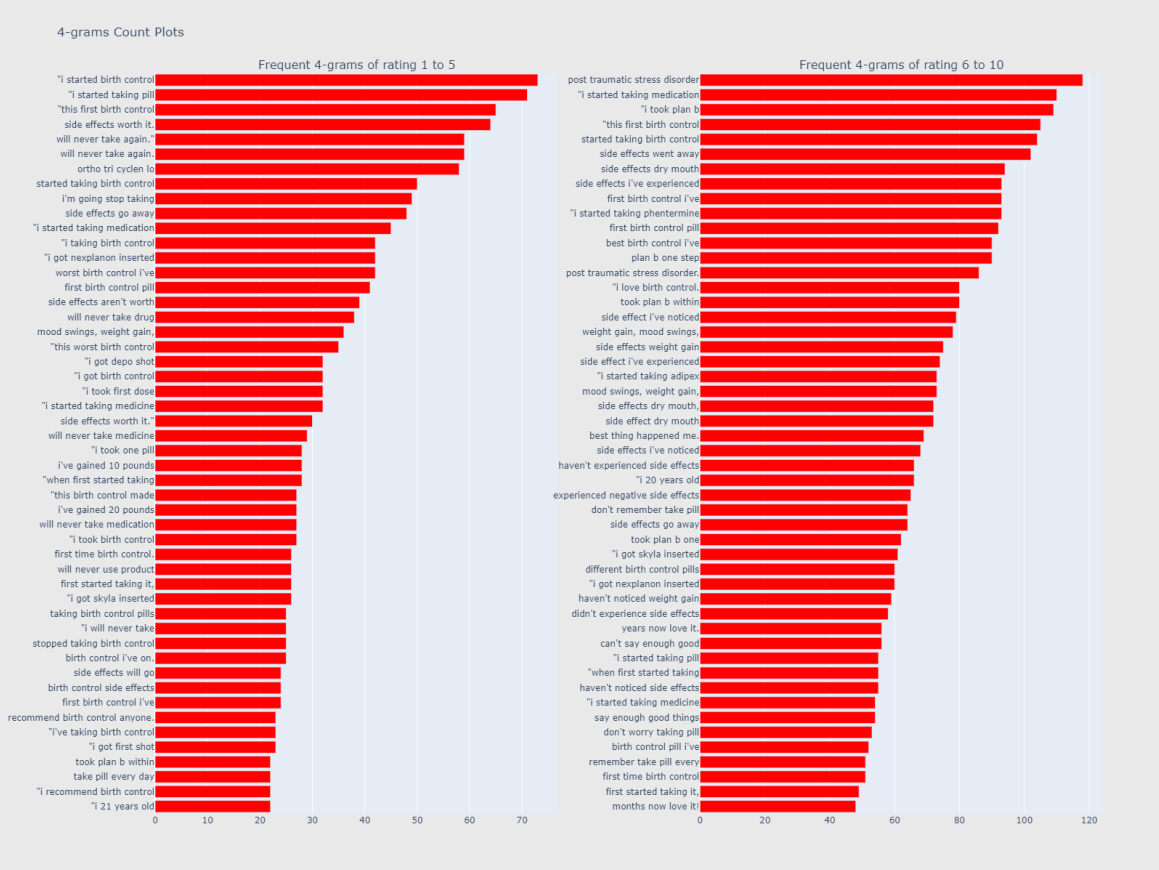


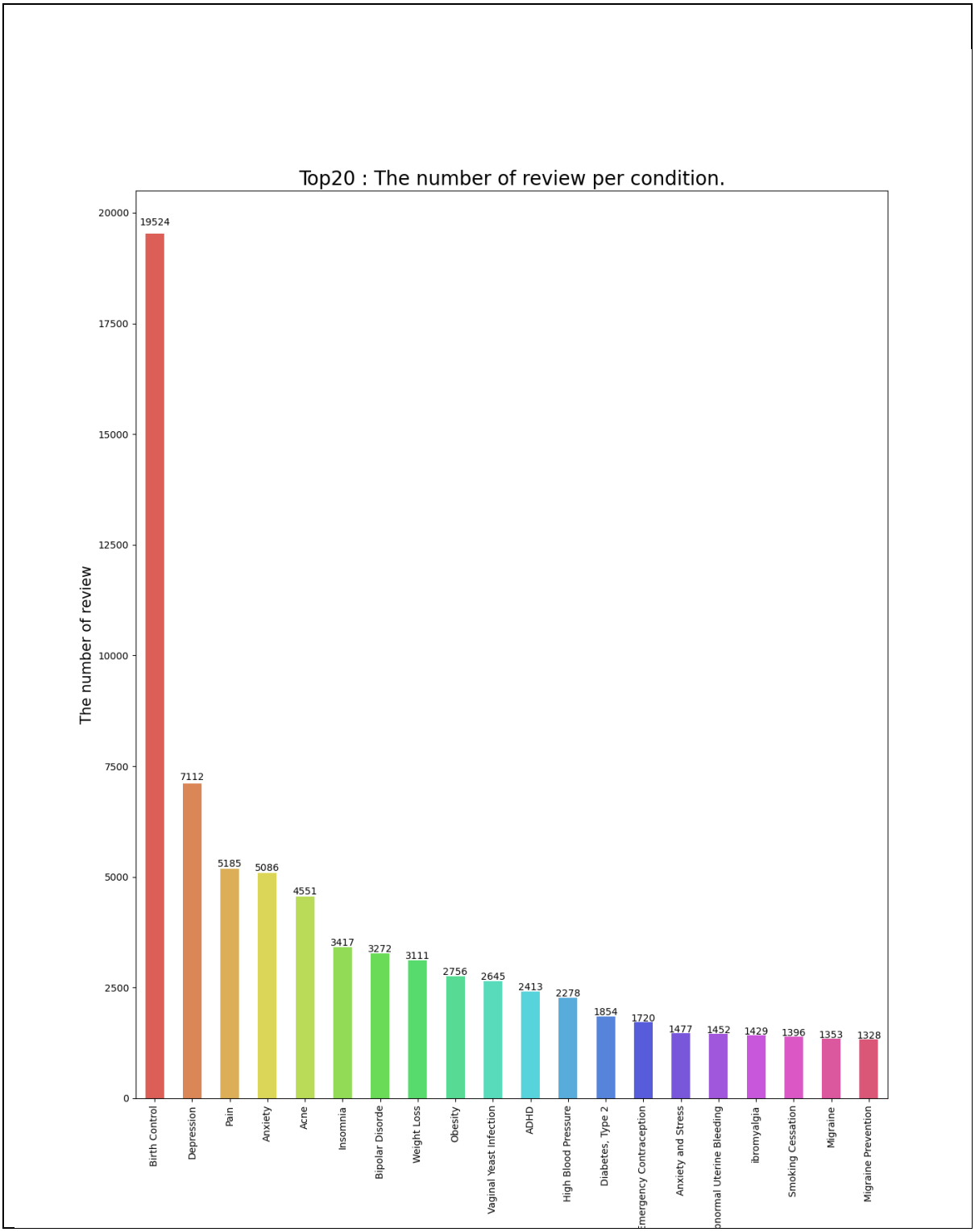


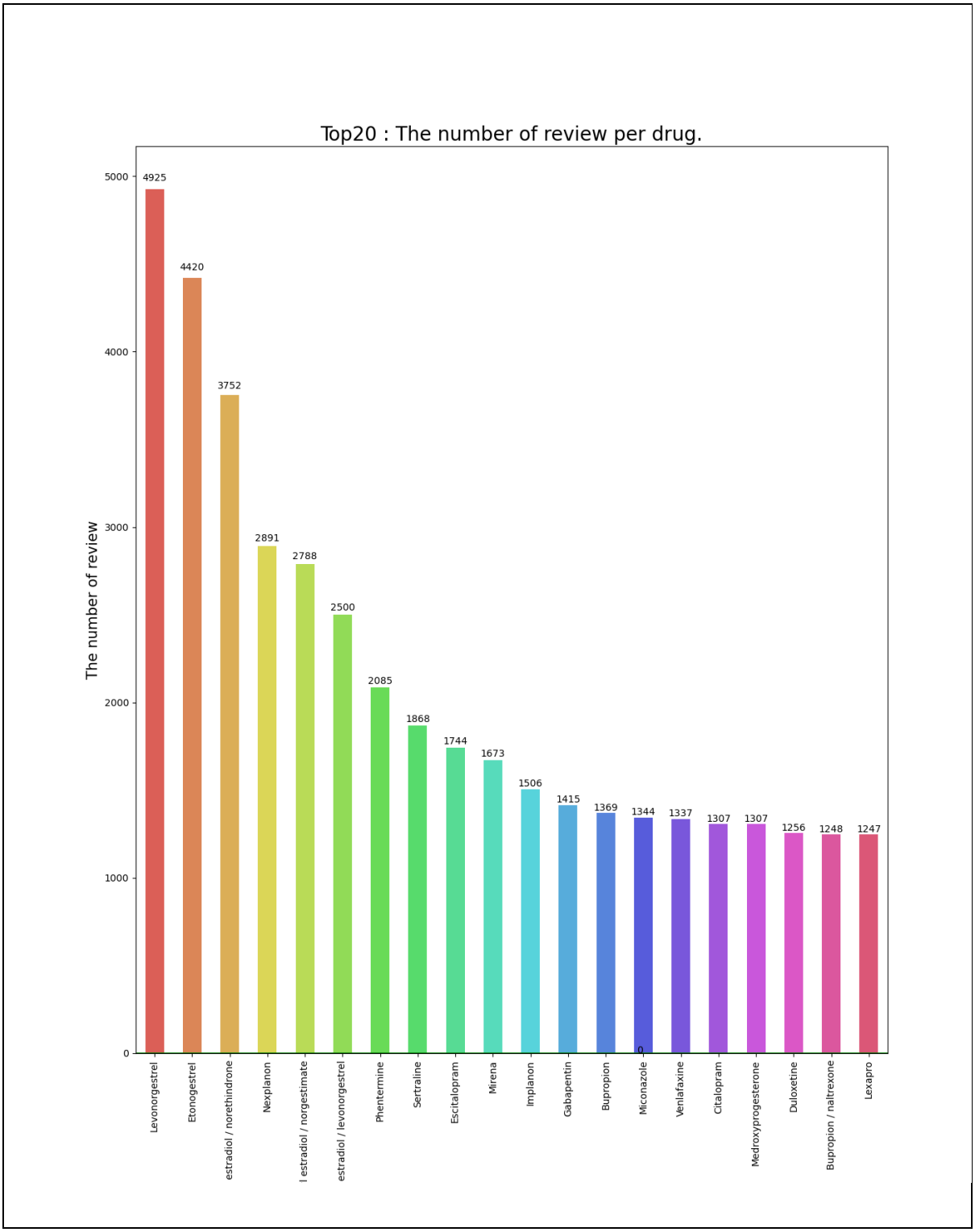


1, and 2- gram shows same distribution both positive(rating>5) and negative(rating<=5) at top 5

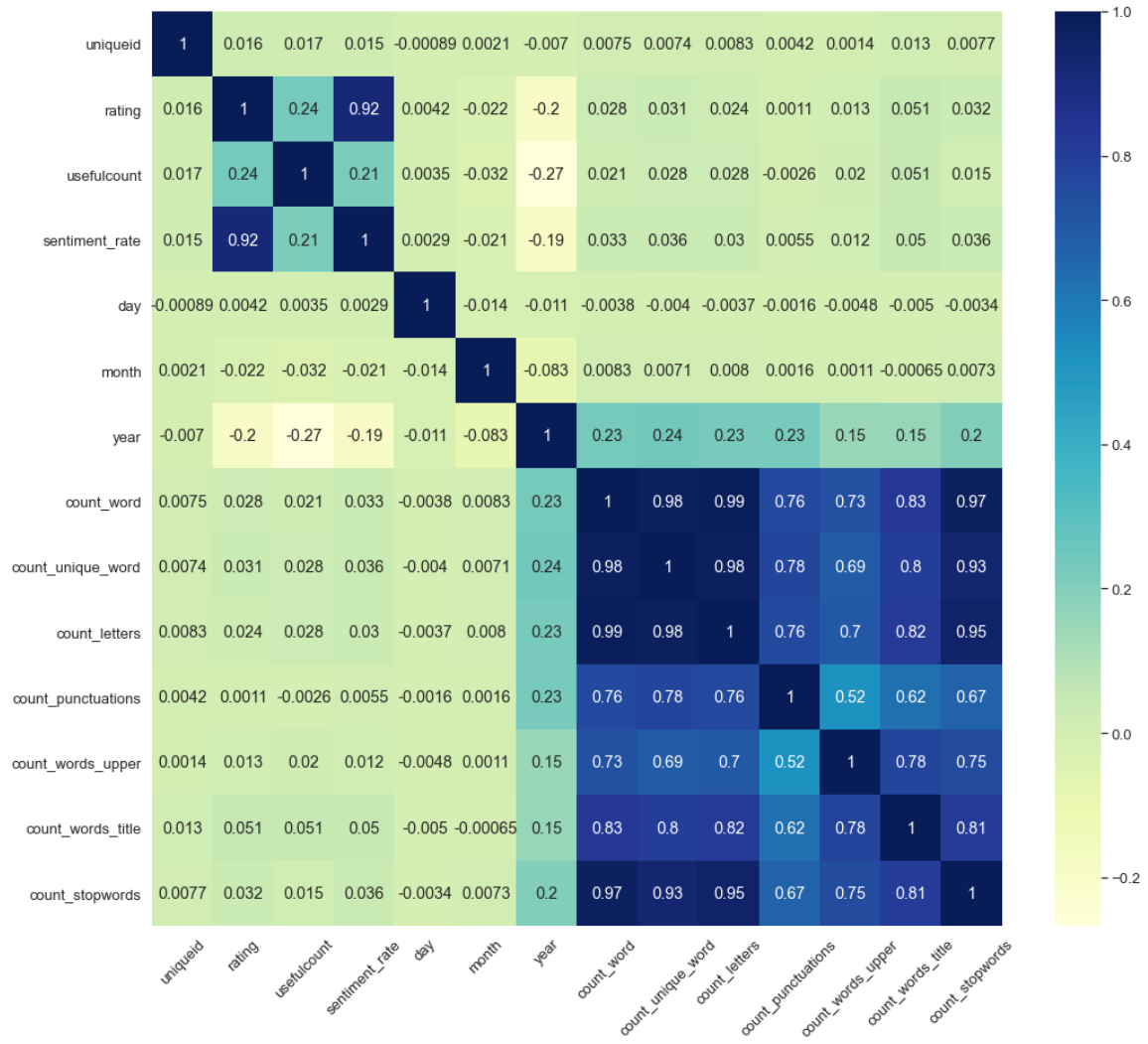



4-grams Count Plots
Frequent 4-grams of rating 1 to 5
"i started birth control
"i started taking pill
"this first birth control
side effects worth it.
will never take again.
will never take again.
ortho tri cyclen lo
started taking birth control
i'm going stop taking
side effects go away
"i started taking medication
"i taking birth control
"i got nexplanon inserted
worst birth control i've
first birth control pill
side effects aren't worth
will never take drug
mood swings, weight gain.
"this worst birth control
"i got depo shot
"i got birth control
"i took first dose
"i started taking medicine
side effects worth it."
will never take medicine
"i took one pill
i've gained 10 pounds
"when first started taking
"this birth control made
i've gained 20 pounds
will never take medication
"i took birth control
first time birth control.
will never use product
first started taking it.
"i got skyla inserted
taking birth control pills
"i will never take
stopped taking birth control
birth control i've on.
side effects will go
birth control side effects
first birth control i've
recommend birth control anyone.
"i've taking birth control
"i got first shot
took plan b within
take pill every day
"i recommend birth control
"i 21 years old
0
10
20
30
40
50
60
70
Frequent 4-grams of rating 6 to 10
post traumatic stress disorder
"i started taking medication
"i took plan b
"this first birth control
started taking birth control
side effects went away
side effects dry mouth
side effects i've experienced
first birth control i've
"i started taking phentermine
first birth control pill
best birth control i've
plan b one step
post traumatic stress disorder.
"i love birth control.
took plan b within
side effect i've noticed
weight gain, mood swings.
side effects weight gain
side effect i've experienced
"i started taking adipex
mood swings, weight gain.
side effects dry mouth.
side effect dry mouth
best thing happened me.
side effects i've noticed
haven't experienced side effects
"i 20 years old
experienced negative side effects
don't remember take pill
side effects go away
took plan b one
"i got skyla inserted
different birth control pills
"i got nexplanon inserted
haven't noticed weight gain
didn't experience side effects
years now love it.
can't say enough good
"i started taking pill
"when first started taking
haven't noticed side effects
"i started taking medicine
say enough good things
don't worry taking pill
birth control pill i've
remember take pill every
first time birth control
first started taking it.
months now love it
0
20
40
60
80
100
120


Top20 : The number of review per condition.
The number of review
0
2500
5000
7500
10000
12500
15000
17500
20000
19524
7112
5185
5086
4551
3417
3272
3111
2756
2645
2413
2278
1854
1720
1477
1452
1429
1396
1353
1328
Birth Control
Depression
Pain
Anxiety
Acne
Insomnia
Bipolar Disorde
Weight Loss
Obesity
Vaginal Yeast Infection
ADHD
High Blood Pressure
Diabetes, Type 2
Emergency Contraception
Anxiety and Stress
onormal Uterine Bleeding
ibromyalgia
Smoking Cessation
Migraine
Migraine Prevention


Top20 : The number of review per drug.
The number of review
5000
4000
3000
2000
1000
0
4925
4420
3752
2891
2788
2500
2085
1868
1744
1673
1506
1415
1369
1344
0
1337
1307
1307
1256
1248
1247
Levonorgestrel
Etonogestrel
estradiol / norethindrone
Nexplanon
l estradiol / norgestimate
estradiol / levonorgestrel
Phentermine
Sertraline
Escitalopram
Mirena
Implanon
Gabapentin
Bupropion
Miconazole
Venlafaxine
Citalopram
Medroxyprogesterone
Duloxetine
Bupropion / naltrexone
Lexapro

Correlation heatmap of the features engineered



데이터셋의 한계점및 생각한 대안

저희 5 명이서 각자 나뉘어 kaggle 에 1 위 부터 5 위 까지 랭크된 포스트를 분석 하였습니다. 그러나 모든 모델이 비슷하게 리뷰를 이용해서 감성(positive, negative)분석을 진행하였고 최종 예측으로 rating 을 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