Observations For EDA

* Similar conditions have different medicines and hence different side effects.
* Doxycycline for example has one review as "1" and other review as "10" but the lowest review is from only one person.So it can be ignored and maximum will be considered.
* Id is unique as per drug and condition and review. Sample submission has id and rating ..need more details, can't it be drug and rating
* Drug names are not unique, for different conditions there are similar drugs
* Input for algo will be drug name and review. A person who is checking the rating already knows the conditions what he has... what's important for that person is the rating of that medicine.
* output is rating
* Review data needs to be cleaned
* **medicineDf.info()**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 40324 entries, 0 to 40323

Data columns (total 7 columns):

Id 40324 non-null int64

drugName 40324 non-null object

condition 40105 non-null object \*\*\*\* This indicates that condition has some null values

review 40324 non-null object

rating 40324 non-null int64

date 40324 non-null object

usefulCount 40324 non-null int64

dtypes: int64(3), object(4)

* **medicineDf.isnull().sum()**

Out[138]:

Id 0

drugName 0

**condition 219 \*\*\*\* 219 Null values**

review 0

rating 0

date 0

usefulCount 0

dtype: int64

* medicineDf.describe()

Out[139]:

Id rating usefulCount

count 40324.000000 40324.000000 40324.000000

mean 20161.500000 6.978896 27.845675

std 11640.680464 3.285145 35.635493

min 0.000000 1.000000 0.000000

25% 10080.750000 4.000000 6.000000

50% 20161.500000 8.000000 16.000000

75% 30242.250000 10.000000 36.000000

max 40323.000000 10.000000 730.000000

* **drugs = pd.value\_counts(medicineDf.drugName)**
* **drugs**

Out[141]:

Levonorgestrel 938

Etonogestrel 813

Ethinyl estradiol / norethindrone 688

Nexplanon 559

Ethinyl estradiol / norgestimate 506

Gentian violet 1

Pulmicort Turbuhaler 1

Aldesleukin 1

Corlanor 1

Onfi 1

Name: drugName, Length: 2419, dtype: int64

* **drugs.head(20) Drugs with maximum reviews**

Out[142]:

**Levonorgestrel 938**

Etonogestrel 813

Ethinyl estradiol / norethindrone 688

Nexplanon 559

Ethinyl estradiol / norgestimate 506

Ethinyl estradiol / levonorgestrel 477

Phentermine 412

Sertraline 374

Escitalopram 343

Mirena 307

Implanon 300

Gabapentin 279

Bupropion 273

Miconazole 253

Citalopram 252

Clonazepam 239

Venlafaxine 237

Duloxetine 236

Varenicline 235

Tramadol 235

Name: drugName, dtype: int64

* **drugs[drugs == drugs.min()].head(20) Drugs with least rating 1**

Out[148]:

Limbitrol DS 1

Betaseron 1

FiberCon 1

Lo / Ovral 1

Fenofibric acid 1

Luxiq 1

Coal tar 1

pHisoHex 1

Glucose 1

Follicle stimulating hormone 1

H.P. Acthar Gel 1

Uroxatral 1

ActoPlus Met 1

Mycophenolic acid 1

Flurandrenolide 1

Axid 1

Dyrenium 1

Chlorcyclizine / phenylephrine 1

Kaletra 1

Pegloticase 1

Name: drugName, dtype: int64

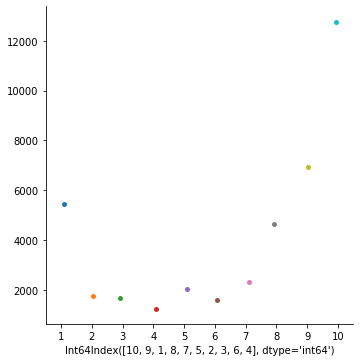
* **Drugs plot**

**sns.countplot(drugs)**

**Out[170]: <matplotlib.axes.\_subplots.AxesSubplot at 0x20f850b5bc8>**



* **sns.catplot(x=rateDf.index,y=rateDf.rating,data=rateDf)**



* **Count as per ratings, maximum rating as 10**

rating

10 12756

9 6916

1 5459

8 4625

7 2298

5 2020

2 1770

3 1652

6 1580

4 1248