Final Report: Analyzing a Mental Health Dataset

Group E - Scarlett Hwang, Bingying Liu, Joaquin Menendez, Nathan Scheperle, Muxin Diao

1) Load all of your data into a Map-Reduce system and set up your tools for data analysis. You'll want to write a basic mapper and a reducer you can use as a starting point.

   Please see the code in appendix.

2) Basic descriptive statistics: How many hospitals are represented in the data? What is the average number of patients per hospital? Minimum and maximum?

   a. Hospital numbers

count(DISTINCT siteId)	**17**

   This answer could vary if depending which table we decide to use.

   b. Maximum

max(num)	**65,443**

   c. Minimum

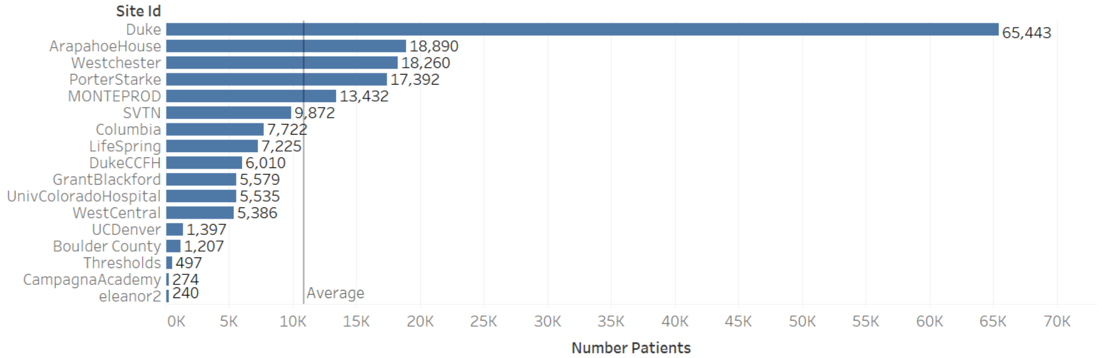
min(num)	**240**

   d. Average

avgNumperHspt	**10,844**

#2. Number of patients per hospital

(Avg 10844 , Max 'Duke' 65443, Min 'eleanor2' 240)



3) Our study has decided to focus on depression and depression-related conditions (Bipolar Disorder, Dysthymic Disorder, etc.). How many of the patients have a depression or depression related diagnosis?

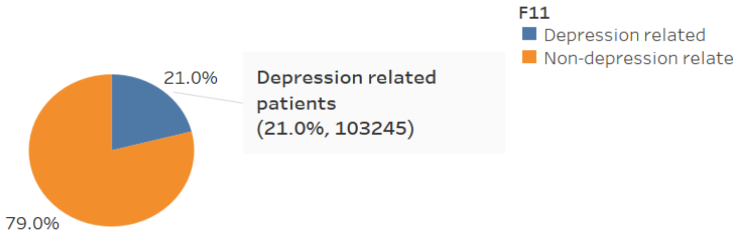
   First, we started looking through some of the different diagnosis and found that there are 1,565 diagnosis. And we also tried to look at top 20 recurrent diagnosis.

   Finally, we chose diagnosis that contained the words: `Bipolar, Dysthymic, Depression, Depressive, Cyclothymic, Cyclothymia` as the diagnoses more related with depression.

   So the answer to the question(total number of depressive related patients) is 103,245.

3. Depression related patients are around 21.0% of total patients

(number of diagnosis related to depression: 103,245, total diagnosis: 388,309)



4) Psych drugs – how many unique ones are in the data (check the "PsyMed" column)?

count(DISTINCT NDC)	943

When we look the amount of different drugs using only the generic drug we observe fewer different types of drugs. Nevertheless, some of the inputs present more than one input (e.g. ASPIRIN; CAFFEINE; SALICYLAMIDE).

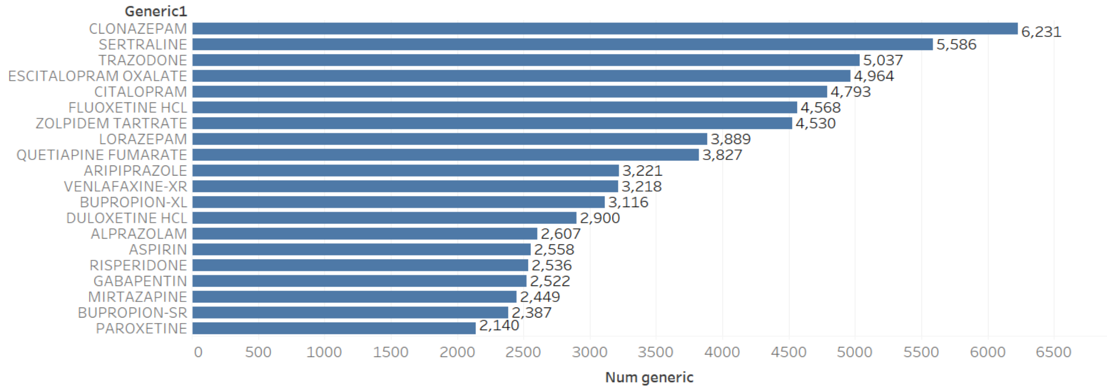
count(DISTINCT Generic)	295

5) Let’s start getting some useful results. What are the most common psych meds for patients with Major Depressive Disorder? For any diagnosis related to depression? What about Cyclothymia?

a. Major Depressive Disorder

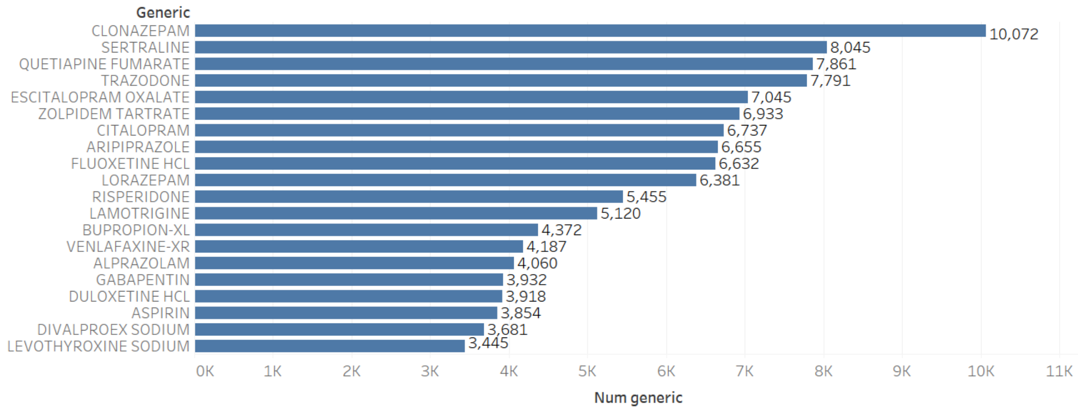
Given the fact that the Diagnosis are not standardized, we decided to treat every diagnosis that has the words Major Depression or Major Depressive as the diagnosis 'Major Depressive Disorder'.

#5-1. Major Depressive disorder



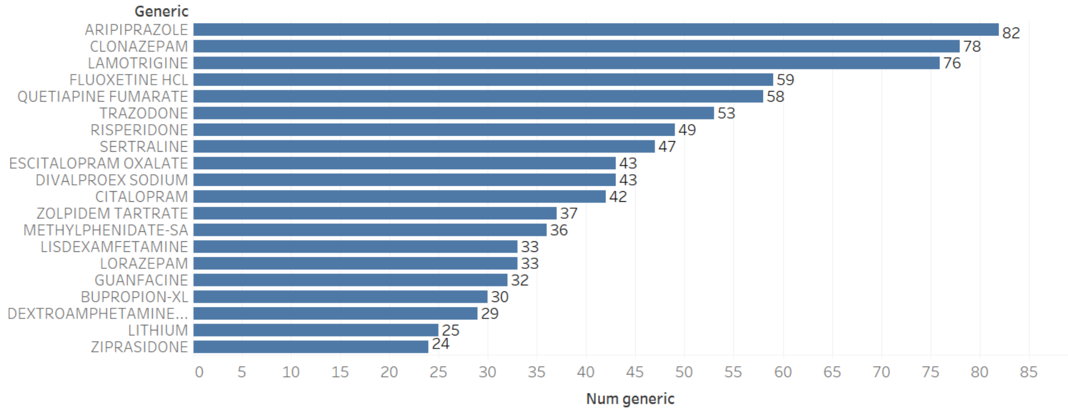
b. All depression related

#5-2. All Depression Related



c. Cyclothymia

#5-3. Cyclothymia



6) Similarly, what are the most common drugs prescribed in patients diagnosed with Bipolar Disorder? Does this vary appreciably by hospital?

Top 3 most commonly prescribed medications for depression by site.

< DEPRESSION >

siteId	Generic	count	rank
LifeSpring	TRAZODONE	312	1
LifeSpring	ARIPIPRAZOLE	265	2
LifeSpring	ESCITALOPRAM OXALATE	243	3
eleanor2	SERTRALINE	8	1
eleanor2	QUETIAPINE FUMARATE	5	2
eleanor2	RISPERIDONE	3	3
eleanor2	ARIPIPRAZOLE	3	3
eleanor2	OLANZAPINE	3	3
ArapahoeHouse	CLORAZEPATE	86	1
ArapahoeHouse	CITALOPRAM	55	2

(showing first 10 rows)

Top 3 most commonly prescribed medications for bipolar by site.
<BIPOLAR>

siteId	Generic	count	rank
LifeSpring	LAMOTRIGINE	291	1
LifeSpring	ARIPIPRAZOLE	270	2
LifeSpring	CLONAZEPAM	252	3
ArapahoeHouse	CLORAZEPATE	48	1
ArapahoeHouse	QUETIAPINE FUMARATE	34	2
ArapahoeHouse	TRAZODONE	34	2
SVTN	QUETIAPINE FUMARATE	400	1
SVTN	ARIPIPRAZOLE	367	2
SVTN	ZOLPIDEM TARTRATE	315	3
DukeCCFH	ARIPIPRAZOLE	16	1

(showing first 10 rows)

7) Is there evidence of a progression of different drugs? In other words, do depression or bipolar patients seem to start out being prescribed certain drugs, and are there drugs that are reserved for cases where the most typical drugs don't work? (Hint: Yes. Yes there are.) What are some of these progressions?

Yes, there is a progression of different drugs for bipolar as well as depressive patients. We calculated average ednum (ednum represent the number clinic visits, thus could be treated as a record of progression) of each drug and ranked the average in ascending order and found that 'LEVOTHYROXINE SODIUM' was the first drug doctor usually prescribed to bipolar patients. And then Lamotrigine, Methadone Hcl, etc. What is quite within expectation is that 'LITHIUM' is almost the last medicine prescribed to bipolar/major depressive patients after any other antidepressants.

8) Drugs often have side effects, sometimes minor and sometimes serious. Are there psych drugs that seem to be prescribed alongside blood pressure medications more often? I may as well warn you – clonidine is used as a blood pressure medication and as a psych med. It does all kinds of things. That one is going to be an outlier.
The table below is the top 20 psych medicines prescribed alongside blood pressure medication.

fluoxetine hcl
nicotine
divalproex sodium-spinkle
hydrocodone bitartrate; acetominophen; alcohol;
memantine
lithium
metformin hcl-xl
liothyronine
carbamazepine
hydrocodone bitartrate; ibuprofen
ginkgo biloba

risperidone m-tab
chlorpromazine hcl-inj
oxycodone
doxepin
clonazepam
clozapine
olanzapine-im
granisetron hcl

9) You have a good-sized collection of data in front of you. Find some interesting patterns. The Clinical Global Impressions (CGI) Scales are used to quickly indicate severity and improvement (since first treatment) of a patient. What inferences can we make about specific medications given the CGI scores? Formulate some hypotheses and test them using the data.

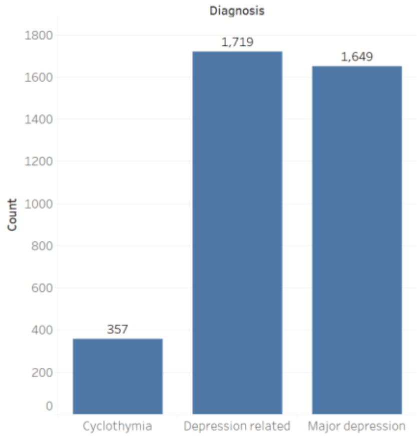
Given answers in 7, we considered that 'Lithium' is only for really severely depressive patients. We used patients' first visit to clinics result to see what severity level patients were, given antidepressants like Lithium, Quetiapine, Aripiprazole and Clonazepam.(filter by 'improvement IS NULL') With Bipolar diagnostic, we expected to see that patients with more severe diagnose were given Lithium more than Aripiprazole or Quetiapine. However, following table for query results showed that this wasn't the case. Lithium was prescribed to less severe patients than Quetiapine, Aripiprazole and Clonazepam. It might be because that Lithium is one of the most widely used and studied mood-stabilizing drug and usually takes several weeks for it to begin working. Once it begins working, it's super efficient and is deemed as the last source to control bipolar disorder. Therefore, doctor prescribe it to patients slightly early on to reduce severity and frequency of mania.

Medicine	**Average Severity for First-visit Patient Prescribed to Certain Medicine**
Lithium	**3.760233918128655**
Quetiapine	**3.928909952606635**
Aripiprazole	**4.393034825870647**
Clonazepam	**3.8737373737373737**

[Appendix]

#5.

#5-4. Number of Different Drugs Prescribed



Code

1)Load all of your data into a Map-Reduce system and set up your tools for data analysis. You’ll want to write a basic mapper and a reducer you can use as a starting point.

```
# To start a spark session and run all these context objects
# you need to write the next line in the command line:
# PYTHONSTARTUP=load_data.py pyspark
from pyspark.sql import SparkSession #importing SQL in order to not need to create temp tables
spark = SparkSession \
	.builder \
	.appName("Easy E - IDS 706 Final Project - Mental Health") \
	.getOrCreate()
Background =
spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_Background.txt')
CGI = spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_CGI.txt')
Meds = spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_Meds.txt')
Patient_Service =
spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_Patient_Service.txt')
PDiagnose =
spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_PDiagnose.txt')
TypePatient =
spark.read.format('csv').option('header','true').load('/shared/mindlinc/VDL2011_TypePatient.txt')
```

2) Basic descriptive statistics: How many hospitals are represented in the data? What is the average number of patients per hospital? Minimum and maximum?

HOW MANY HOSPITALS

```
CGI.select(CGI['siteId']).distinct().count()
17
```

Or the SQL version

```
>>> CGI.createOrReplaceTempView("tempCGI")
>>> sqlCGI = spark.sql('SELECT COUNT( DISTINCT siteId) FROM tempCGI')
>>> sqlCGI.show()
+----------------------+
|count(DISTINCT siteId)|
+----------------------+
|                    17|
+----------------------+
```

This answer could vary if depending wich table we decide to use. If we use the Background table to adress this we would enter the following query:

>>> Background.createOrReplaceTempView("Backtemp")
>>> spark.sql('SELECT siteId, count(distinct ID) AS Number_Patients FROM Backtemp GROUP BY siteId ORDER BY Number_Patients').show()

```
+--------------------+---------------+
|              siteId|Number_Patients|
+--------------------+---------------+
|            eleanor2|            240|
|     CampagnaAcademy|            274|
|          Thresholds|            497|
|      Boulder County|           1207|
|            UCDenver|           1397|
|         WestCentral|           5386|
|UnivColoradoHospital|           5535|
|      GrantBlackford|           5579|
|            DukeCCFH|           6010|
|          LifeSpring|           7225|
|            Columbia|           7722|
|                SVTN|           9872|
|           MONTEPROD|          13432|
|        PorterStarke|          17392|
|         Westchester|          18260|
|       ArapahoeHouse|          18890|
|                Duke|          65443|
+--------------------+---------------+
```

MAXIMUM
spark.sql('SELECT siteId,MAX(site.num) FROM (SELECT COUNT(*) AS num FROM Backtemp GROUP BY siteId) AS site').show()

```
+--------+
|max(num)|
+--------+
|   65443|
+--------+
```

MINIMUM
>>> sqlCGI = spark.sql('SELECT MIN(site.num) FROM (SELECT COUNT(*) AS num FROM Backtemp GROUP BY siteId) AS site')
>>> sqlCGI.show()

```
+--------+
|min(num)|
+--------+
|     240|
+--------+
```

AVERAGE AMOUNT OF NUMBER OF PATIENTS PER HOSPITAL #Given 'ID' SHOULD BE the Key for every hospital is not going to be repeated inside hospital (siteId)
>>> Background.select(Background['ID']).count() / Background.select(Background['siteId']).distinct().count()
10844

3) Our study has decided to focus on depression and depression-related conditions (Bipolar Disorder, Dysthymic Disorder, etc.). How many of the patients have a depression or depression related diagnosis?

First we started looking trough some of the different Diagnosis.
>>> PDiagnose.createOrReplaceTempView('tempPDiagnose')

```
>>> diagnosis = spark.sql('SELECT COUNT(distinct diagnosis) AS N_DIAGNOSTICS FROM (SELECT DISTINCT site Id, BackgroundID, Diagnosis from tempPDiagnose)')
>>> diagnosis.show()
+-------------+
|N_DIAGNOSTICS|
+-------------+
|         1565|
+-------------+
```

Top recurrent diagnosis entries (It could include different entries for same patients)

```
diagnosis = spark.sql('SELECT DISTINCT diagnosis, count (*) as amount from tempPDiagnose group by diagnosis order by amount')
diagnosis.show(20, False)
+-----------------------------------------------------------------+-------+
|diagnosis                                                        |amount |
+-----------------------------------------------------------------+-------+
|Diagnosis Deferred on Axis II                                    |1100257|
|No Diagnosis on Axis II                                          |382461 |
|Post-Traumatic Stress Disorder                                   |373862 |
|Schizoaffective Disorder                                         |341126 |
|Schizophrenia Paranoid Type                                      |254937 |
|Alcohol Dependence                                               |237101 |
|ADHD Combined                                                    |222535 |
|Major Depressive Disorder Recurrent Moderate                     |215927 |
|Anxiety Disorder NOS                                             |200831 |
|Major Depressive Disorder Recurrent Severe without Psychotic Features|192930 |
|Depressive Disorder NOS                                          |191110 |
|Oppositional Defiant Disorder                                    |188886 |
|Mood Disorder NOS                                                |180856 |
|Alcohol Abuse                                                    |180854 |
|Generalized Anxiety Disorder                                     |180704 |
|Bipolar Disorder NOS                                             |171534 |
|Borderline Personality Disorder                                  |169743 |
|Cannabis Abuse                                                   |138792 |
|Dysthymic Disorder                                               |124450 |
|Opioid Dependence                                                |107278 |
+-----------------------------------------------------------------+-------+
only showing top 20 rows
```

We realize that there a lot of different diagnosis. Looking trough them we chose diagnosis that contained the words: `Bipolar, Dysthymic, Depression, Depressive, Cyclothymic, Cyclothymia` as the diagnoses more related with depresion.

```
>>> sqlPDiagnose = spark.sql('SELECT DISTINCT siteId, BackgroundID, Diagnosis from tempPDiagnose')
>>> diagnosis = sqlPDiagnose.toPandas()
>>> diagnosis['Depression diagnosis'] = 0 #to create a column with empty values
>>> import re
>>> diagnosis['Depression diagnosis'][diagnosis.Diagnosis.str.contains('Bipolar|Dysthymic|Depression|Depressive|Cyclothymia|Cyclothymic', regex =True, case=False)==True] = 1
```

An example of the dataframe

```
>>> diagnosis[diagnosis['Depression diagnosis'] == 1].head(10)
        siteId BackgroundID                                     Diagnosis  Depression diagnosis
89    ArapahoeHouse        11303                       Depressive Disorder NOS                     1
126   ArapahoeHouse        15592   Bipolar I Disorder Single Episode Manic Mild                     1
```

```
147  CampagnaAcademy      69                 Bipolar Disorder II           1
155       Columbia       538                Dysthymic Disorder            1
157       Columbia       667            Depressive Disorder NOS           1
163       Columbia       959            Depressive Disorder NOS           1
164       Columbia       980               Bipolar Disorder II            1
165       Columbia      1040              Bipolar Disorder NOS            1
166       Columbia      1054  Major Depressive Disorder Single Episode Sever...           1
170       Columbia      1264               Bipolar Disorder II            1
>>>
```

Total depressive related patients

```
>>> sum(diagnosis['Depression diagnosis'])
103245
```

Total other diagnosis

```
>>> diagnosis.shape[0] - sum(diagnosis['Depression diagnosis'])
388309
```

Proportion of Depressive over Total

```
>>> sum(diagnosis['Depression diagnosis'])/float(diagnosis.shape[0])
0.21003796124128785
```

4) Psych drugs – how many unique ones are in the data (check the "PsyMed" column)?

In this case we assume that a Psych drug is unique using the National Drug Code (NDC). This code allow us to distinguish a drug base on the manufacturer,dosage form and package size. We used this variable instead of Medication or Generic because this variable where filled in a very inconsistend ways.

```
Meds.createOrReplaceTempView("Medtemp")
sqlMed = spark.sql("SELECT DISTINCT NDC FROM Medtemp WHERE PsyMed = 'Yes' ")
sqlMed = spark.sql("SELECT COUNT(DISTINCT  NDC) FROM Medtemp WHERE PsyMed = 'Yes' ")
sqlMed.show()
+-------------------+
|count(DISTINCT NDC)|
+-------------------+
|                943|
```

When we look the amount of different drugs using only the generic drug we observe fewer different types of drugs. Nevertheless, some of the inputs present more than one input (e.g. ASPIRIN; CAFFEINE; SALICYLAMIDE).

```
>>> sqlMed = spark.sql("SELECT COUNT(DISTINCT  Generic) FROM Medtemp WHERE PsyMed = 'Yes' ")
>>> sqlMed.show()
+-----------------------+
|count(DISTINCT Generic)|
+-----------------------+
|                    295|
+-----------------------+
```

5) Let's start getting some useful results. What are the most common psych meds for patients with Major Depressive Disorder? For any diagnosis related to depression? What about Cyclothymia?

Given the fact that the Diagnosis are not standarized we decided to treat every diagnosis that has the words Major Depression or Major Depressive as the diagnosis 'Major Depressive Disorder'

```
>>> Background.createOrReplaceTempView('Backgroundtemp')
```

```
>>> PDiagnose.createOrReplaceTempView("PDtemp")
>>> Meds.createOrReplaceTempView("Medtemp")

>>> atemptable = spark.sql("SELECT Back.*,PD.Diagnosis FROM Backgroundtemp as Back  LEFT
JOIN PDtemp as PD ON Back.ID = PD.BackgroundID AND Back.siteID = PD.siteID \
                WHERE PD.Diagnosis LIKE '%ajor %epression%' OR PD.Diagnosis LIKE
'%ajor %epressive%'")
>>> atemptable.createOrReplaceTempView('Jointemp')

>>> reduced = spark.sql('SELECT DISTINCT siteId, ID, Diagnosis FROM Jointemp GROUP BY
Diagnosis, siteId, Id')
>>> reduced.createOrReplaceTempView('Jointemp')

>>> sqlJoin = spark.sql("SELECT JJ.*, Medtemp.Generic FROM Jointemp AS JJ LEFT JOIN Medtemp
ON (Medtemp.BackgroundID = JJ.ID) AND (Medtemp.siteId = JJ.siteId)")
>>> sqlJoin.createOrReplaceTempView('Jointemp2')

>>> dep_med = spark.sql('SELECT distinct generic,siteid, id, count( distinct id) as NUM FROM
Jointemp2 group by siteid,id,generic order by id desc')
>>> dep_med.createOrReplaceTempView('dep_temp')

>>> spark.sql('SELECT generic, count(generic) as Num_generic from dep_temp group by generic order
by Num_generic DESC').show()


+--------------------+-----------+
|             generic|Num_generic|
+--------------------+-----------+
|          CLONAZEPAM|       6231|
|          SERTRALINE|       5586|
|           TRAZODONE|       5037|
|ESCITALOPRAM OXALATE|       4964|
|          CITALOPRAM|       4793|
|      FLUOXETINE HCL|       4568|
|   ZOLPIDEM TARTRATE|       4530|
|           LORAZEPAM|       3889|
| QUETIAPINE FUMARATE|       3827|
|         ARIPIPRAZOLE|       3221|
|      VENLAFAXINE-XR|       3218|
|        BUPROPION-XL|       3116|
|      DULOXETINE HCL|       2900|
|          ALPRAZOLAM|       2607|
|             ASPIRIN|       2558|
|         RISPERIDONE|       2536|
|          GABAPENTIN|       2522|
|         MIRTAZAPINE|       2449|
|        BUPROPION-SR|       2387|
|          PAROXETINE|       2140|
+--------------------+-----------+
only showing top 20 rows


+--------+
|count   |
+--------+
|    1649|
+--------+
```

We could see that we have 1649 different types of Generic Drugs for Mayor Depression.

We proceed in similar way to the Diagnosis 'Cyclothymia'.

```
>>> spark.sql('SELECT DISTINCT Diagnosis FROM PDtemp WHERE Diagnosis LIKE
"%yclothy%"').show(10,False)
+------------------------+
|Diagnosis               |
+------------------------+
|Cyclothymic Disorder    |
|Cyclothymic Disorder  R/O|
|Cyclothymic Disorders   |
|r/o cyclothymia         |
+------------------------+
```

We observe that there are similar labels to refer to the same diagnosis.

```
>> sqlJoin1 = spark.sql("SELECT Back.siteId, Back.ID,PD.Diagnosis FROM Backgroundtemp as Back
LEFT JOIN PDtemp as PD ON Back.ID = PD.BackgroundID AND Back.siteID = PD.siteID WHERE
PD.Diagnosis  LIKE '%yclothy%'")
>> sqlJoin1.createOrReplaceTempView('Jointemp1')
>> sqlJoin2 = spark.sql("SELECT JT.*, Medtemp.Generic FROM Jointemp1 AS JT LEFT JOIN
Medtemp ON (Medtemp.BackgroundID = JT.ID) AND (Medtemp.siteId = JT.siteId)")
>> sqlJoin2.createOrReplaceTempView('Jointemp2')
>> cyc_med = spark.sql('SELECT distinct generic,siteid, id, count( distinct id) as NUM FROM Jointemp2
group by siteid,id,generic order by id desc')
>> cyc_med.createOrReplaceTempView('cyc_temp')
>> spark.sql('SELECT generic, count(generic) as Num_generic from cyc_temp group by generic order by
Num_generic DESC').show()


+--------------------+-----------+
|             generic|Num_generic|
+--------------------+-----------+
|        ARIPIPRAZOLE|         82|
|          CLONAZEPAM|         78|
|         LAMOTRIGINE|         76|
|      FLUOXETINE HCL|         59|
| QUETIAPINE FUMARATE|         58|
|           TRAZODONE|         53|
|         RISPERIDONE|         49|
|          SERTRALINE|         47|
|ESCITALOPRAM OXALATE|         43|
|   DIVALPROEX SODIUM|         43|
|          CITALOPRAM|         42|
|   ZOLPIDEM TARTRATE|         37|
|  METHYLPHENIDATE-SA|         36|
|           LORAZEPAM|         33|
|    LISDEXAMFETAMINE|         33|
|          GUANFACINE|         32|
|        BUPROPION-XL|         30|
|DEXTROAMPHETAMINE...|         29|
|             LITHIUM|         25|
|         ZIPRASIDONE|         24|
+--------------------+-----------+

>>> sqlJoin3.createOrReplaceTempView("cyc_num")
>>> spark.sql('SELECT COUNT(*) FROM  cyc_num').show()
```

```
+--------+
|count(1)|
+--------+
|     357|
+--------+
```

We proceed in similar way to the 'Depresion related' Diagnoses

```
>>> sqlJoin1 = spark.sql("SELECT Back.siteId, Back.ID,PD.Diagnosis FROM Backgroundtemp as Back LEFT JOIN PDtemp as PD ON Back.ID = PD.BackgroundID AND Back.siteID = PD.siteID WHERE UPPER(PD.Diagnosis) LIKE '%DEPRESSIVE%' OR UPPER(PD.Diagnosis) LIKE '%CYCLOTHY%' OR  UPPER(PD.Diagnosis) LIKE '%DEPRESSION%' OR UPPER(PD.Diagnosis) LIKE '%BIPOLAR%' OR PD.Diagnosis LIKE '%DYSTHYMI%' ")

>>> sqlJoin1.createOrReplaceTempView('Jointemp1')
>>> sqlJoin2 = spark.sql("SELECT JT.*, Medtemp.Generic FROM Jointemp1 AS JT LEFT JOIN Medtemp ON (Medtemp.BackgroundID = JT.ID) AND (Medtemp.siteId = JT.siteId)")
>>> sqlJoin2.createOrReplaceTempView('Jointemp2')

>>> dep_rel_med = spark.sql('SELECT distinct generic,siteid, id, count( distinct id) as NUM FROM Jointemp2 group by siteid,id,generic order by id desc')
>>> dep_rel_med.createOrReplaceTempView('dep_rel_temp')
>>> sqlJoin3 = spark.sql('SELECT generic, count(generic) as Num_generic from dep_rel_temp group by generic order by Num_generic DESC')
>>> sqlJoin3.createOrReplaceTempView("dep_rel_temp")
>>> sqlJoin3.show()


+--------------------+-----------+
|             generic|Num_generic|
+--------------------+-----------+
|          CLONAZEPAM|      10072|
|          SERTRALINE|       8045|
| QUETIAPINE FUMARATE|       7861|
|           TRAZODONE|       7791|
|ESCITALOPRAM OXALATE|       7045|
|   ZOLPIDEM TARTRATE|       6933|
|          CITALOPRAM|       6737|
|        ARIPIPRAZOLE|       6655|
|       FLUOXETINE HCL|      6632|
|           LORAZEPAM|       6381|
|         RISPERIDONE|       5455|
|         LAMOTRIGINE|       5120|
|        BUPROPION-XL|       4372|
|       VENLAFAXINE-XR|      4187|
|          ALPRAZOLAM|       4060|
|          GABAPENTIN|       3932|
|      DULOXETINE HCL|       3918|
|             ASPIRIN|       3854|
|   DIVALPROEX SODIUM|       3681|
|LEVOTHYROXINE SODIUM|       3445|
+--------------------+-----------+
only showing top 20 rows

>>> spark.sql('SELECT COUNT(*) FROM  dep_rel_temp').show()
+--------+
|count(1)|
+--------+
```

```
|    1818|
+--------+
```

6) Similarly, what are the most common drugs prescribed in patients diagnosed with Bipolar Disorder? Does this vary appreciably by hospital? Top 3 most commonly prescribed medications for depression by site.

*****DEPRESSION*****

```
>>> site_common_depress.select('*', f.rank().over(window).alias('rank')).filter(f.col('rank') <= 3).show(51, False)
+--------------------+--------------------+-----+----+
|siteId              |Generic             |count|rank|
+--------------------+--------------------+-----+----+
|LifeSpring          |TRAZODONE           |312  |1   |
|LifeSpring          |ARIPIPRAZOLE        |265  |2   |
|LifeSpring          |ESCITALOPRAM OXALATE|243  |3   |
|eleanor2            |SERTRALINE          |8    |1   |
|eleanor2            |QUETIAPINE FUMARATE |5    |2   |
|eleanor2            |RISPERIDONE         |3    |3   |
|eleanor2            |ARIPIPRAZOLE        |3    |3   |
|eleanor2            |OLANZAPINE          |3    |3   |
|ArapahoeHouse       |CLORAZEPATE         |86   |1   |
|ArapahoeHouse       |CITALOPRAM          |55   |2   |
|ArapahoeHouse       |FLUOXETINE HCL      |52   |3   |
|SVTN                |ZOLPIDEM TARTRATE   |1028 |1   |
|SVTN                |QUETIAPINE FUMARATE |688  |2   |
|SVTN                |ESCITALOPRAM OXALATE|676  |3   |
|DukeCCFH            |SERTRALINE          |41   |1   |
|DukeCCFH            |FLUOXETINE HCL      |39   |2   |
|DukeCCFH            |METHYLPHENIDATE-SA  |33   |3   |
|GrantBlackford      |TRAZODONE           |493  |1   |
|GrantBlackford      |CITALOPRAM          |405  |2   |
|GrantBlackford      |CLONAZEPAM          |277  |3   |
|Duke                |CLONAZEPAM          |3175 |1   |
|Duke                |SERTRALINE          |2984 |2   |
|Duke                |LORAZEPAM           |2149 |3   |
|UCDenver            |CLONAZEPAM          |102  |1   |
|UCDenver            |SERTRALINE          |96   |2   |
|UCDenver            |ZOLPIDEM TARTRATE   |91   |3   |
|UnivColoradoHospital|SERTRALINE          |425  |1   |
|UnivColoradoHospital|CITALOPRAM          |411  |2   |
|UnivColoradoHospital|CLONAZEPAM          |396  |3   |
|MONTEPROD           |SERTRALINE          |724  |1   |
|MONTEPROD           |ZOLPIDEM TARTRATE   |686  |2   |
|MONTEPROD           |QUETIAPINE FUMARATE |649  |3   |
|Thresholds          |SERTRALINE          |1    |1   |
|Thresholds          |ESCITALOPRAM OXALATE|1    |1   |
|CampagnaAcademy     |ARIPIPRAZOLE        |8    |1   |
|CampagnaAcademy     |TRAZODONE           |6    |2   |
|CampagnaAcademy     |ZIPRASIDONE         |6    |2   |
|Westchester         |ESCITALOPRAM OXALATE|867  |1   |
|Westchester         |CLONAZEPAM          |858  |2   |
|Westchester         |QUETIAPINE FUMARATE |845  |3   |
|PorterStarke        |TRAZODONE           |1041 |1   |
|PorterStarke        |ARIPIPRAZOLE        |752  |2   |
|PorterStarke        |ESCITALOPRAM OXALATE|684  |3   |
|WestCentral         |CLONAZEPAM          |103  |1   |
|WestCentral         |TRAZODONE           |98   |2   |
|WestCentral         |CITALOPRAM          |88   |3   |
```

```
|Columbia           |CLONAZEPAM         |556  |1   |
|Columbia           |CITALOPRAM         |360  |2   |
|Columbia           |LORAZEPAM          |334  |3   |
|Boulder County     |DULOXETINE HCL     |4    |1   |
|Boulder County     |TRAZODONE          |2    |2   |
+--------------------+--------------------+-----+----+
```

Top 3 most commonly prescribed medications for bipolar by site.

```
>>> site_common.select('*', f.rank().over(window).alias('rank')).filter(f.col('rank') <=
3).filter(f.col('count') >= 10).show(51, False)
+--------------------+-------------------+-----+----+
|siteId              |Generic            |count|rank|
+--------------------+-------------------+-----+----+
|LifeSpring          |LAMOTRIGINE        |291  |1   |
|LifeSpring          |ARIPIPRAZOLE       |270  |2   |
|LifeSpring          |CLONAZEPAM         |252  |3   |
|ArapahoeHouse       |CLORAZEPATE        |48   |1   |
|ArapahoeHouse       |QUETIAPINE FUMARATE|34   |2   |
|ArapahoeHouse       |TRAZODONE          |34   |2   |
|SVTN                |QUETIAPINE FUMARATE|400  |1   |
|SVTN                |ARIPIPRAZOLE       |367  |2   |
|SVTN                |ZOLPIDEM TARTRATE  |315  |3   |
|DukeCCFH            |ARIPIPRAZOLE       |16   |1   |
|DukeCCFH            |SERTRALINE         |11   |2   |
|GrantBlackford      |TRAZODONE          |160  |1   |
|GrantBlackford      |ARIPIPRAZOLE       |127  |2   |
|GrantBlackford      |CLONAZEPAM         |113  |3   |
|Duke                |CLONAZEPAM         |1098 |1   |
|Duke                |LAMOTRIGINE        |997  |2   |
|Duke                |DIVALPROEX SODIUM  |966  |3   |
|UCDenver            |LAMOTRIGINE        |130  |1   |
|UCDenver            |QUETIAPINE FUMARATE|112  |2   |
|UCDenver            |LITHIUM CARBONATE  |79   |3   |
|UnivColoradoHospital|QUETIAPINE FUMARATE|341  |1   |
|UnivColoradoHospital|LAMOTRIGINE        |282  |2   |
|UnivColoradoHospital|CLONAZEPAM         |234  |3   |
|MONTEPROD           |QUETIAPINE FUMARATE|606  |1   |
|MONTEPROD           |ARIPIPRAZOLE       |503  |2   |
|MONTEPROD           |RISPERIDONE        |466  |3   |
|Westchester         |QUETIAPINE FUMARATE|801  |1   |
|Westchester         |ARIPIPRAZOLE       |711  |2   |
|Westchester         |DIVALPROEX SODIUM  |645  |3   |
|PorterStarke        |ARIPIPRAZOLE       |757  |1   |
|PorterStarke        |LAMOTRIGINE        |660  |2   |
|PorterStarke        |TRAZODONE          |625  |3   |
|WestCentral         |LAMOTRIGINE        |30   |1   |
|WestCentral         |CLONAZEPAM         |29   |2   |
|WestCentral         |QUETIAPINE FUMARATE|28   |3   |
|Columbia            |CLONAZEPAM         |287  |1   |
|Columbia            |LAMOTRIGINE        |238  |2   |
|Columbia            |QUETIAPINE FUMARATE|179  |3   |
+--------------------+-------------------+-----+----+
```

```
# Query for 6
>>> query = "select distinct m.siteId, m.BackgroundID, m.Medication, upper(m.Generic) as Generic,
m.ednum \
from sqlMeds m \
inner join (select siteId, BackgroundID from sqlPDiagnose where lower(Diagnosis) like '%bipolar%') p \
```

```
on m.siteId = p.siteId and m.BackgroundID = p.BackgroundID \
where m.PsyMed = 'yes' "
# and trim(Generic) in (select trim(Generic) from sqlBipolar) "
# Get the medications for students diagnosed with bipolar
>>> bipolar = sql(query).persist()

# Get the distinct generic drugs these patients are prescribed
>>> patient_drugs = bipolar.select("siteId", "BackgroundID", "Generic").distinct()
# Calculate the most common generic drugs for bipolar patients by site
>>> site_common = patient_drugs.groupBy("siteId", "Generic").count().persist()
# Calculate the most commonly prescribed drugs to all bipolar patients
>>> most_common = patient_drugs.groupBy("Generic").count().orderBy("count",
ascending=False).limit(20).persist()
>>> most_common.createOrReplaceTempView("sqlBipolar")
# Window for partitioning by site
>>> window =Window.partitionBy(site_common['siteId']).orderBy(site_common['count'].desc())
# Calculate the most cmmonly prescribed drugs to bipolar patients by site
>>> site_common.select('*', f.rank().over(window).alias('rank')).filter(f.col('rank') <=
3).filter(f.col('count') >= 10).show(51, False)
# Convert ednum to numeric
>>> bipolar = bipolar.withColumn('ednum',f.col('ednum').cast('integer'))
# For each drug a patient is prescribed, keep the row for the first visit where they were prescribed it
>>> bipolar_distinct = bipolar.groupBy('siteId', 'BackgroundID',
'Generic').min('ednum').withColumnRenamed('min(ednum)', 'ednum')

#bipolar.groupBy('Generic').count().orderBy('count', ascending=False).show(20, False)
# For each bipolar patient, sort the drugs they have been prescribed by order of visit
>>> sorted_bipolar = ( bipolar_distinct.alias('a').join(most_common.alias('b'), f.col('a.Generic') ==
f.col('b.Generic'), 'inner').groupBy('siteId', 'BackgroundID')
                    .agg(f.sort_array( f.collect_list( f.struct( f.col('ednum'), f.col('a.Generic') ) ), asc = True)
                    .alias('sorted_meds') )
   )
>>> sorted_bipolar.show(50,False)
>>> drug_order = bipolar_distinct.select('siteId','BackgroundID','Generic',
f.rank().over(Window.partitionBy("siteId",'BackgroundID').orderBy("ednum")).alias("rank"))
drug_order.groupBy('Generic').agg({'rank':'mean', '*':'count'}).filter(f.col('count(1)') >=
75).orderBy('avg(rank)').show(50, False)
query = "select distinct p.Generic \
from (select siteId, BackgroundID, Generic from sqlMeds where PsyMed = 'yes') p \
inner join (select siteId, BackgroundID from sqlMeds where bpMeds = '1') b \
on trim(p.BackgroundID) = trim(b.BackgroundID) and trim(p.siteId) = trim(b.siteId) "
>>> psych_meds_side_effect = sql(query).persist()
>>> psych_meds_side_effect.show()
# Repeating 6 for depression
>>> query = "select distinct p.siteId, m.BackgroundID, m.Medication, upper(m.Generic) as Generic,
m.ednum \
from sqlMeds m \
inner join (select siteId, BackgroundID from sqlPDiagnose where lower(Diagnosis) like '%depres%') p \
on m.siteId = lower(p.siteId) and m.BackgroundID = p.BackgroundID \
where m.PsyMed = 'yes' "
# and trim(Generic) in (select trim(Generic) from sqlBipolar) "
# Get the medications for students diagnosed with bipolar
depress = sql(query).persist()

# Get the distinct generic drugs these patients are prescribed
>>> patient_drugs_depress = depress.select("siteId", "BackgroundID", "Generic").distinct()
# Calculate the most common generic drugs for bipolar patients by site
>>> site_common_depress = patient_drugs_depress.groupBy("siteId", "Generic").count().persist()
```

```
# Calculate the most commonly prescribed drugs to all bipolar patients
>>> most_common_depress = patient_drugs_depress.groupBy("Generic").count().orderBy("count", ascending=False).limit(20).persist()
# Window for partitioning by site
>>> window = Window.partitionBy(site_common_depress['siteId']).orderBy(site_common_depress['count'].desc())
# Calculate the most cmmonly prescribed drugs to bipolar patients by site
>>> site_common_depress.select('*', f.rank().over(window).alias('rank')).filter(f.col('rank') <= 3).show(51, False)
```

7)Is there evidence of a progression of different drugs? In other words, do depression or bipolar patients seem to start out being prescribed certain drugs, and are there drugs that are reserved for cases where the most typical drugs don't work? (Hint: Yes. Yes there are.) What are some of these progressions?

```
>>> bipolar = bipolar.withColumn('ednum',f.col('ednum').cast('integer'))

>>> bipolar_distinct = bipolar.groupBy('siteId', 'BackgroundID', 'Generic').min('ednum').withColumnRenamed('min(ednum)', 'ednum')

>>> drug_order = bipolar_distinct.select('siteId','BackgroundID','Generic', f.rank().over(Window.partitionBy("siteId",'BackgroundID').orderBy("ednum")).alias("rank"))

>>> drug_order.groupBy('Generic').agg({'rank':'mean', '*':'count'}).filter(f.col('count(1)') >= 75).orderBy('avg(rank)').show(50, False)
+--------------------------------+------------------+--------+
|Generic                         |avg(rank)         |count(1)|
+--------------------------------+------------------+--------+
|LEVOTHYROXINE SODIUM            |1.7620751341681575|1118    |
|LAMOTRIGINE                     |2.2815506508206   |3534    |
|METHADONE HCL                   |2.383211678832117 |274     |
|CLORAZEPATE                     |2.4533333333333333 |75      |
|CLONAZEPAM                      |2.45629466739967  |3638    |
|DIVALPROEX SODIUM               |2.4712945590994373|2665    |
|SERTRALINE                      |2.5172018348623855|1744    |
|BUPRENOPHINE HCL; NALOXONE HCL-SL|2.5376884422110555|199     |
|QUETIAPINE FUMARATE             |2.579866092778575 |4182    |
|FLUOXETINE HCL                  |2.6039087947882735|1535    |
|ARIPIPRAZOLE                    |2.6488267861850776|3793    |
|LITHIUM CARBONATE               |2.6681574239713775|2236    |
|LISDEXAMFETAMINE                |2.6933333333333334|600     |
|ALPRAZOLAM                      |2.6953316953316953|1221    |
|ESCITALOPRAM OXALATE            |2.7309236947791167|1743    |
|RISPERIDONE                     |2.740052063964299 |2689    |
|CLONIDINE                       |2.7417417417417416|333     |
|CITALOPRAM                      |2.7745222929936304|1570    |
|METHYLPHENIDATE-SA              |2.785607196401799 |667     |
|DEXTROAMPHETAMINE; AMPHETAMINE-XR|2.79343365253078  |731     |
|DIVALPROEX SODIUM-ER            |2.8396501457725947|1372    |
|DULOXETINE HCL                  |2.839779005524862 |1086    |
|PAROXETINE                      |2.882608695652174 |690     |
|TRAZODONE                       |2.897741273100616 |2435    |
|DEXMETHYLPHENIDATE HCL-XR       |2.9017094017094016|234     |
|GABAPENTIN                      |2.9264            |1250    |
|LORAZEPAM                       |2.937471051412691 |2159    |
|GUANFACINE                      |2.9444444444444446|306     |
|TOPIRAMATE                      |2.9825986078886313|862     |
|OXCARBAZEPINE                   |2.9947848761408085|767     |
|OLANZAPINE                      |3.0834834834834837|1665    |
```

```
|OXYCODONE                       |3.131487889273356 |289     |
|ZOLPIDEM TARTRATE               |3.1472129585516915|2099    |
|BENZTROPINE                     |3.1847389558232932|996     |
|MELATONIN                       |3.1983471074380163|121     |
|DONEPEZIL                       |3.2061855670103094|97      |
|HALOPERIDOL LACTATE             |3.2665006226650064|803     |
|ZIPRASIDONE                     |3.2693333333333334|1125    |
|FLUVOXAMINE MALEATE             |3.269503546099291 |141     |
|DIAZEPAM                        |3.291015625       |512     |
|VENLAFAXINE-XR                  |3.295973884657236 |919     |
|HALOPERIDOL DECANOATE           |3.3               |120     |
|CARBAMAZEPINE                   |3.332688588007737 |517     |
|RISPERIDONE M-TAB               |3.3380952380952382|210     |
|QUETIAPINE FUMARATE XR          |3.348155156102176 |1057    |
|BUPROPION-XL                    |3.3583415597235935|1013    |
|TRAMADOL                        |3.400709219858156 |282     |
|ATOMOXETINE HCL                 |3.4047619047619047|420     |
|LITHIUM                         |3.4292237442922375|876     |
|BUSPIRONE                       |3.4814814814814814|540     |
+--------------------------------+------------------+--------+
only showing top 50 rows
```

8) Drugs often have side effects, sometimes minor and sometimes serious. Are there psych drugs that seem to be prescribed alongside blood pressure medications more often? I may as well warn you – clonidine is used as a blood pressure medication and as a psych med. It does all kinds of things. That one is going to be an outlier.

```
>>> query = "select distinct p.Generic \
... from (select siteId, BackgroundID, Generic from sqlMeds where PsyMed = 'yes') p \
... inner join (select siteId, BackgroundID from sqlMeds where bpMeds = '1') b \
... on trim(p.BackgroundID) = trim(b.BackgroundID) and trim(p.siteId) = trim(b.siteId) "
>>> psych_meds_side_effect = sql(query).persist()
>>> psych_meds_side_effect.show(20, False)
+---------------------------------------------+
|Generic                                      |
+---------------------------------------------+
|fluoxetine hcl                               |
|nicotine                                     |
|divalproex sodium-spinkle                    |
|cannot be determined                         |
|hydrocodone bitartrate; acetominophen; alcohol;|
|memantine                                    |
|lithium                                      |
|metformin hcl-xl                             |
|liothyronine                                 |
|carbamazepine                                |
|hydrocodone bitartrate; ibuprofen            |
|ginkgo biloba                                |
|risperidone m-tab                            |
|chlorpromazine hcl-inj                       |
|oxycodone                                    |
|doxepin                                      |
|clonazepam                                   |
|clozapine                                    |
|olanzapine-im                                |
|granisetron hcl                              |
+---------------------------------------------+
only showing top 20 rows
```

9) You have a good-sized collection of data in front of you. Find some interesting patterns. The Clinical Global Impressions (CGI) Scales are used to quickly indicate severity and improvement (since first treatment) of a patient. What inferences can we make about specific medications given the CGI scores? Formulate some hypotheses and test them using the data.

```
sqlJoin1 = spark.sql("SELECT P.Diagnosis, P.siteId, P.BackgroundID,cgi.severity,cgi.improvement, cgi.ednum FROM PDiagnose as p INNER JOIN CGI ON P.BackgroundID = CGI.BackgroundID AND \
                    P.siteID = CGI.siteID AND P.ednum = CGI.ednum WHERE UPPER(P.Diagnosis) LIKE '%BIPOLAR%'").persist()
sqlJoin1.createOrReplaceTempView('sqlJoin1')
sqlJoin2 = spark.sql("SELECT C.*, M.generic FROM sqlJoin1 as C inner join Meds as M on C.siteId = M.siteId AND c.BackgroundID = m.BackgroundID AND m.ednum = c.ednum").persist()
aripiprazole = sqlJoin2.select('*').where('generic LIKE "%ARIPIPRAZOLE%" and improvement IS NULL AND SEVERITY IS NOT NULL').persist()
quetiapine = sqlJoin2.select('*').where('generic LIKE "%QUETIAPINE%" and improvement IS NULL AND SEVERITY IS NOT NULL').persist()
lithium = sqlJoin2.select('*').where('generic LIKE "%LITHIUM%" and improvement IS NULL AND SEVERITY IS NOT NULL').persist()
clonazepam = sqlJoin2.select('*').where('generic LIKE "%CLONAZEPAM%" and improvement IS NULL AND SEVERITY IS NOT NULL').persist()
ZIPRASIDONE = sqlJoin2.select('*').where('generic LIKE "%ZIPRASIDONE%" and improvement IS NULL AND SEVERITY IS NOT NULL').persist()
#Let's see the severity level of Lithium, Quetiapine, Apripiprazole and Clonazepam.
lithium.createOrReplaceTempView('lithium')
quetiapine.createOrReplaceTempView('quetiapine')
aripiprazole.createOrReplaceTempView('aripiprazole')
clonazepam.createOrReplaceTempView('clonazepam')
spark.sql('Select avg(severity) as Severity_Lithium from lithium').show()
spark.sql('Select avg(severity) as Severity_Quetiapine from quetiapine').show()
spark.sql('Select avg(severity) as Severity_Aripiprazole from aripiprazole').show()
spark.sql('Select avg(severity) as Severity_Clonazepam from clonazepam').show()
clonazepan.show()
spark.sql('Select avg(severity) as Severity_Ziprasidone from ZIPRASIDONE').show()
ZIPRASIDONE.show()
```

```
+-----------------+
| Severity_Lithium|
+-----------------+
|3.760233918128655|
+-----------------+
+-------------------+
|Severity_Quetiapine|
+-------------------+
|  3.928909952606635|
+-------------------+
+---------------------+
|Severity_Aripiprazole|
+---------------------+
|    4.393034825870647|
+---------------------+
+-------------------+
|Severity_Clonazepam|
+-------------------+
| 3.8737373737373737|
+-------------------+
+--------------------+------------+------------+--------+-----------+------+----------+
|           Diagnosis|      siteId|BackgroundID|severity|improvement| ednum|   generic|
+--------------------+------------+------------+--------+-----------+------+----------+
```

Bipolar Disorder NOS	Duke	14812	4	null	532189	CLONAZEPAM
Bipolar I Most Re...	Duke	17225	5	null	448659	CLONAZEPAM
Bipolar I Disorde...	Duke	50212	4	null	385863	CLONAZEPAM
Bipolar Disorder II	Duke	50212	4	null	385863	CLONAZEPAM
Bipolar I Most Re...	Duke	17225	4	null	282923	CLONAZEPAM
Bipolar I Most Re...	Duke	17225	4	null	413200	CLONAZEPAM
Bipolar I Most Re...	Duke	17225	3	null	559883	CLONAZEPAM
Bipolar I Most Re...	Duke	25789	3	null	127981	CLONAZEPAM
Bipolar I Most Re...	Duke	30970	3	null	337377	CLONAZEPAM
Bipolar I Disorde...	Duke	50212	3	null	538366	CLONAZEPAM
Bipolar Disorder II	Duke	50212	3	null	538366	CLONAZEPAM
Bipolar Disorder II	Duke	52642	5	null	22913	CLONAZEPAM
Bipolar I Most Re...	Duke	8027	4	null	3314	CLONAZEPAM
Bipolar I Disorde...	SVTN	5384	3	null	78546	CLONAZEPAM
Bipolar I Disorde...	SVTN	5384	3	null	78546	CLONAZEPAM
Bipolar I Disorde...	MONTEPROD	4317	4	null	75909	CLONAZEPAM
Bipolar I Disorde...	PorterStarke	10118	5	null	325108	CLONAZEPAM
Bipolar I Most Re...	Duke	25789	3	null	147636	CLONAZEPAM
Bipolar I Most Re...	Duke	52165	4	null	96072	CLONAZEPAM
Bipolar Disorder II	Duke	42223	4	null	119253	CLONAZEPAM

Severity_Ziprasidone
4.1891891891891895

Diagnosis	siteId	BackgroundID	severity	improvement	ednum	generic
Bipolar I Most Re...	Duke	17225	4	null	413200	ZIPRASIDONE
Bipolar I Disorde...	Duke	47038	4	null	541707	ZIPRASIDONE
Bipolar Disorder II	LifeSpring	4898	3	null	246041	ZIPRASIDONE
Bipolar I Most Re...	Duke	17225	4	null	518576	ZIPRASIDONE
Bipolar I Disorde...	PorterStarke	8009	5	null	447580	ZIPRASIDONE
Bipolar I Disorde...	Duke	36479	2	null	295959	ZIPRASIDONE
Bipolar I Disorde...	LifeSpring	2278	6	null	18472	ZIPRASIDONE
Bipolar I Disorde...	SVTN	8117	3	null	101945	ZIPRASIDONE
Bipolar Disorder II	PorterStarke	15111	6	null	442767	ZIPRASIDONE
Bipolar I Disorde...	SVTN	6990	4	null	366910	ZIPRASIDONE
Bipolar Disorder NOS	MONTEPROD	11802	4	null	237462	ZIPRASIDONE
Bipolar Disorder NOS	PorterStarke	8289	5	null	234	ZIPRASIDONE
Bipolar Disorder II	LifeSpring	4898	3	null	195087	ZIPRASIDONE
Bipolar I Most Re...	Duke	17225	4	null	398770	ZIPRASIDONE
Bipolar Disorder NOS	PorterStarke	8289	6	null	279338	ZIPRASIDONE
Bipolar I Disorde...	SVTN	6990	4	null	159858	ZIPRASIDONE
Bipolar I Disorde...	SVTN	6990	4	null	309883	ZIPRASIDONE
Bipolar Disorder NOS	Duke	64074	6	null	363462	ZIPRASIDONE
Bipolar I Disorde...	SVTN	6990	4	null	171397	ZIPRASIDONE
Bipolar I Most Re...	Duke	17225	4	null	31480	ZIPRASIDONE