```
>>> df.select(split(col("email"),'@')[1].alias("domain"),"count").groupBy("domai
n").count().orderBy("count",ascending=False).show(50)
+--------------------+-----+

|              domain|count|
+--------------------+-----+
|           gmail.com| 5849|
|          apache.org|  281|
|          redhat.com|  203|
|             163.com|  131|
|         hotmail.com|  119|
|    googlegroups.com|   95|
|             gnu.org|   90|
|          debian.org|   82|
|           yahoo.com|   81|
|         freebsd.org|   78|
|              qq.com|   73|
|          google.com|   72|
|         outlook.com|   72|
|            intel.com|   56|
|               gmx.de|   48|
|        mirantis.com|   35|
|          oracle.com|   33|
|               web.de|   29|
|      protonmail.com|   28|
|      googlemail.com|   27|
|    openjdk.java.net|   27|
|             gmx.net|   25|
|             126.com|   24|
|          python.org|   24|
|           yandex.ru|   23|
|       linux.ibm.com|   22|
|          huawei.com|   22|
|    lists.debian.org|   21|
|       ververica.com|   21|
|             free.fr|   20|
|         t-online.de|   19|
|             arm.com|   19|
|          comcast.net|   19|
|       canonical.com|   18|
|          icloud.com|   18|
|           pobox.com|   18|
|     linux.intel.com|   17|
|             gmx.com|   17|
|           posteo.de|   17|
|            live.com|   16|
|          golang.org|   16|
|          kernel.org|   16|
|syzkaller.appspot...|   15|
|          vmware.com|   15|
|        ericsson.com|   15|
|           nokia.com|   14|
|         foxmail.com|   14|
|           cisco.com|   14|
|          riseup.net|   14|
|          linaro.org|   14|
+--------------------+-----+
only showing top 50 rows

>>> df.select(reverse(split(col("email"),'\.'))[0].alias("TLD"),"count").groupBy
```

```
("TLD").count().orderBy("count",ascending=False).show(50)
+----+-----+
| TLD|count|
+----+-----+
| com| 9287|
| org| 1538|
| net|  535|
|  de|  487|
| edu|  178|
|  uk|  136|
|  fr|  115|
|  io|  112|
|  ca|   81|
|  it|   80|
|  ru|   67|
|  ch|   65|
|  au|   62|
|  nl|   56|
|  at|   50|
|  eu|   49|
|  se|   43|
|  in|   42|
|  jp|   41|
|  pl|   38|
|  us|   38|
|  cn|   38|
|  cz|   31|
|  br|   30|
|  nz|   25|
|  ai|   23|
|  be|   23|
|name|   22|
|  es|   21|
|  me|   20|
|  co|   20|
|  fi|   18|
|  dk|   18|
|  no|   17|
| gov|   15|
|info|   14|
|  cc|   13|
|  gr|   11|
|  ie|   11|
| biz|   10|
| dev|   10|
|  sk|   10|
|  tr|    9|
| xyz|    9|
|  hu|    9|
|  mx|    8|
|  fm|    8|
|  pk|    7|
|  pt|    7|
|  kr|    7|
+----+-----+
only showing top 50 rows

>>> df.count()
13871
```

```
>>> df2.select("*").groupBy("word").count().orderBy("count",ascending=False).sho
w(100)
+-------------+-----+

|         word|count|
+-------------+-----+
|         send|21835|
|      message|19172|
|  unsubscribe|16515|
|         2021|15340|
|         list|14130|
|      mailing|13025|
|         file|12323|
|         mail|12052|
|          jan|11759|
|        email|10633|
|        flink|10238|
|           pm| 9855|
|        group| 9496|
|         code| 9482|
|      problem| 9458|
|         data| 9291|
|     received| 8714|
|         2022| 8621|
|       return| 8505|
|         2020| 8409|
|   discussion| 8390|
|       google| 8365|
|         time| 8275|
|       groups| 8185|
|        visit| 8158|
|   subscribed| 8132|
|         view| 8128|
|        error| 7987|
|          web| 7979|
|    receiving| 7858|
|      version| 7839|
|          set| 7697|
|       emails| 7615|
|       100644| 7330|
|          add| 7267|
|          job| 7188|
|         diff| 7140|
|      running| 7117|
|         info| 6861|
|         case| 6836|
|          dec| 6643|
|         work| 6599|
|      address| 6493|
|        files| 6250|
|         type| 6210|
|         user| 5877|
|       server| 5763|
|         test| 5705|
|         read| 5675|
|       struct| 5648|
|       change| 5590|
|      support| 5496|
|        issue| 5463|
|       python| 5366|
```

```
|     function| 5321|
|        check| 5197|
|          int| 5146|
|       memory| 4937|
|           00| 4855|
|       apache| 4745|
|        start| 4708|
|     messages| 4706|
|       create| 4702|
|         good| 4589|
|          nov| 4583|
|       cygwin| 4576|
|      request| 4455|
|         long| 4404|
|      command| 4222|
|      release| 4210|
|       source| 4193|
|      default| 4179|
|      package| 4085|
|     specific| 3847|
|       static| 3809|
|       device| 3698|
|          log| 3682|
|       failed| 3682|
|         void| 3680|
|       number| 3676|
|configuration| 3670|
|        build| 3514|
|      provide| 3446|
|   connection| 3429|
|        write| 3427|
|      working| 3372|
|       output| 3350|
|      current| 3341|
|        class| 3297|
|       client| 3270|
|          bit| 3260|
|        table| 3232|
|        works| 3189|
|       update| 3187|
|        guest| 3159|
|      process| 3151|
|          lot| 3112|
|         port| 3111|
|      windows| 3103|
|          kvm| 3086|
+-------------+-----+
only showing top 100 rows

>>> df2.count()
2855539
```