4_data_wrangling

July 10, 2021

1 Data Wrangling with Spark

This is the code used in the previous screencast. Run each code cell to understand what the code does and how it works.

These first three cells import libraries, instantiate a SparkSession, and then read in the data set

```
In [4]: from pyspark.sql import SparkSession
        from pyspark.sql.functions import udf
        from pyspark.sql.types import StringType
        from pyspark.sql.types import IntegerType
        from pyspark.sql.functions import desc
        from pyspark.sql.functions import asc
        from pyspark.sql.functions import sum as Fsum

        import datetime

        import numpy as np
        import pandas as pd
        %matplotlib inline
        import matplotlib.pyplot as plt
```

```
In [5]: spark = SparkSession \
            .builder \
            .appName("Wrangling Data") \
            .getOrCreate()
```

```
In [6]: path = "data/sparkify_log_small.json"
        user_log = spark.read.json(path)
```

2 Data Exploration

The next cells explore the data set.

```
In [7]: user_log.take(5)
```

```
Out[7]: [Row(artist='Showaddywaddy', auth='Logged In', firstName='Kenneth', gender='M', itemInSe
         Row(artist='Lily Allen', auth='Logged In', firstName='Elizabeth', gender='F', itemInSes
```

```
Row(artist='Cobra Starship Featuring Leighton Meester', auth='Logged In', firstName='Ve
Row(artist='Alex Smoke', auth='Logged In', firstName='Sophee', gender='F', itemInSessio
Row(artist=None, auth='Logged In', firstName='Jordyn', gender='F', itemInSession=0, las
```

In [8]:
```
user_log.printSchema()
```

```
root
 |-- artist: string (nullable = true)
 |-- auth: string (nullable = true)
 |-- firstName: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- itemInSession: long (nullable = true)
 |-- lastName: string (nullable = true)
 |-- length: double (nullable = true)
 |-- level: string (nullable = true)
 |-- location: string (nullable = true)
 |-- method: string (nullable = true)
 |-- page: string (nullable = true)
 |-- registration: long (nullable = true)
 |-- sessionId: long (nullable = true)
 |-- song: string (nullable = true)
 |-- status: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- userAgent: string (nullable = true)
 |-- userId: string (nullable = true)
```

In [9]:
```
user_log.describe().show()
```

```
+-------+-----------------+----------+---------+------+-----------------+--------+------------
|summary|           artist|      auth|firstName|gender|    itemInSession|lastName|          le
+-------+-----------------+----------+---------+------+-----------------+--------+------------
|  count|             8347|     10000|     9664|  9664|            10000|    9664|
|   mean|            461.0|      null|     null|  null|          19.6734|    null|249.648658749
| stddev|            300.0|      null|     null|  null|25.382114916132597|    null|95.0043713078
|    min|              !!!|     Guest|   Aakash|     F|                0| Acevedo|         1.1
|    max|ÃČÂŞlafur Arnalds|Logged Out|     Zoie|     M|              163|  Zuniga|        1806.
+-------+-----------------+----------+---------+------+-----------------+--------+------------
```

In [10]:
```
user_log.describe("artist").show()
```

```
+-------+-----------------+
|summary|           artist|
+-------+-----------------+
|  count|             8347|
|   mean|            461.0|
```

```
| stddev|            300.0|
|    min|              !!!|
|    max|ÃČÂŞlafur Arnalds|
+-------+-----------------+
```

```
In [11]: user_log.describe("sessionId").show()
```

```
+-------+------------------+
|summary|         sessionId|
+-------+------------------+
|  count|             10000|
|   mean|         4436.7511|
| stddev|2043.1281541827557|
|    min|                 9|
|    max|              7144|
+-------+------------------+
```

```
In [12]: user_log.count()
```

```
Out[12]: 10000
```

```
In [13]: user_log.select("page").dropDuplicates().sort("page").show()
```

```
+----------------+
|            page|
+----------------+
|           About|
|       Downgrade|
|           Error|
|            Help|
|            Home|
|           Login|
|          Logout|
|        NextSong|
|   Save Settings|
|        Settings|
|Submit Downgrade|
|  Submit Upgrade|
|         Upgrade|
+----------------+
```

```
In [14]: user_log.select(["userId", "firstname", "page", "song"]).where(user_log.userId == "1046
```

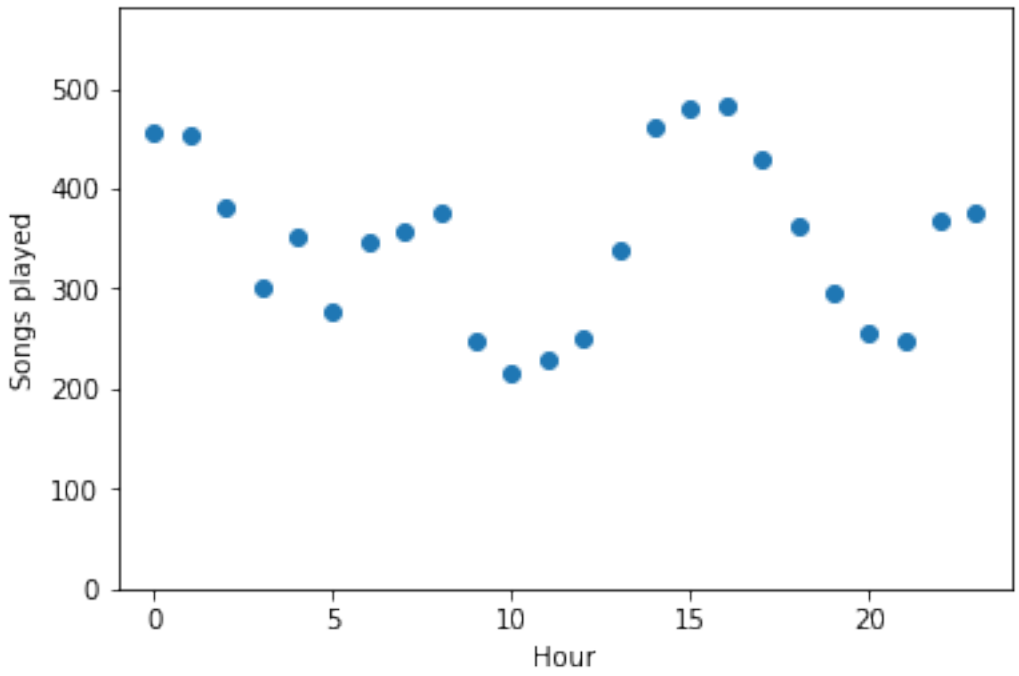
```
Out[14]: [Row(userId='1046', firstname='Kenneth', page='NextSong', song='Christmas Tears Will Fa
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Be Wary Of A Woman'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Public Enemy No.1'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Reign Of The Tyrants'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Father And Son'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='No. 5'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Seventeen'),
          Row(userId='1046', firstname='Kenneth', page='Home', song=None),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='War on war'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Killermont Street'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Black & Blue'),
          Row(userId='1046', firstname='Kenneth', page='Logout', song=None),
          Row(userId='1046', firstname='Kenneth', page='Home', song=None),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Heads Will Roll'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Bleed It Out [Live At M
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Clocks'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Love Rain'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song="Ry Ry's Song (Album Ver
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='The Invisible Man'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Catch You Baby (Steve P
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Ask The Mountains'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Given Up (Album Version
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='El Cuatrero'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Hero/Heroine'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Spring'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Rising Moon'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Tough Little Boys'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song="Qu'Est-Ce Que T'Es Bell
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Secrets'),
          Row(userId='1046', firstname='Kenneth', page='NextSong', song='Under The Gun')]
```

3 Calculating Statistics by Hour

```
In [15]: get_hour = udf(lambda x: datetime.datetime.fromtimestamp(x / 1000.0). hour)

In [16]: user_log = user_log.withColumn("hour", get_hour(user_log.ts))

In [17]: user_log.head()

Out[17]: Row(artist='Showaddywaddy', auth='Logged In', firstName='Kenneth', gender='M', itemInSe

In [18]: songs_in_hour = user_log.filter(user_log.page == "NextSong").groupby(user_log.hour).cou

In [19]: songs_in_hour.show()

+----+-----+
|hour|count|
+----+-----+
```

```
|   0|  456|
|   1|  454|
|   2|  382|
|   3|  302|
|   4|  352|
|   5|  276|
|   6|  348|
|   7|  358|
|   8|  375|
|   9|  249|
|  10|  216|
|  11|  228|
|  12|  251|
|  13|  339|
|  14|  462|
|  15|  479|
|  16|  484|
|  17|  430|
|  18|  362|
|  19|  295|
+----+-----+
only showing top 20 rows
```

```
In [20]: songs_in_hour_pd = songs_in_hour.toPandas()
         songs_in_hour_pd.hour = pd.to_numeric(songs_in_hour_pd.hour)

In [21]: plt.scatter(songs_in_hour_pd["hour"], songs_in_hour_pd["count"])
         plt.xlim(-1, 24);
         plt.ylim(0, 1.2 * max(songs_in_hour_pd["count"]))
         plt.xlabel("Hour")
         plt.ylabel("Songs played");
```



4 Drop Rows with Missing Values

As you'll see, it turns out there are no missing values in the userID or session columns. But there are userID values that are empty strings.

```
In [22]: user_log_valid = user_log.dropna(how = "any", subset = ["userId", "sessionId"])

In [23]: user_log_valid.count()

Out[23]: 10000

In [24]: user_log.select("userId").dropDuplicates().sort("userId").show()

+------+
|userId|
+------+
|      |
|    10|
|   100|
|  1000|
|  1003|
|  1005|
|  1006|
|  1017|
```

```
|  1019|
|  1020|
|  1022|
|  1025|
|  1030|
|  1035|
|  1037|
|   104|
|  1040|
|  1042|
|  1043|
|  1046|
+------+
only showing top 20 rows
```

```
In [25]: user_log_valid = user_log_valid.filter(user_log_valid["userId"] != "")
```

```
In [26]: user_log_valid.count()
```

```
Out[26]: 9664
```

5 Users Downgrade Their Accounts

Find when users downgrade their accounts and then flag those log entries. Then use a window function and cumulative sum to distinguish each user's data as either pre or post downgrade events.

```
In [27]: user_log_valid.filter("page = 'Submit Downgrade'").show()
```

```
+------+---------+---------+------+-------------+--------+------+-----+--------------------+----
|artist|     auth|firstName|gender|itemInSession|lastName|length|level|            location|meth
+------+---------+---------+------+-------------+--------+------+-----+--------------------+----
|  null|Logged In|    Kelly|     F|           24|  Newton|  null| paid|Houston-The Woodl...|   P
+------+---------+---------+------+-------------+--------+------+-----+--------------------+----
```

```
In [28]: user_log.select(["userId", "firstname", "page", "level", "song"]).where(user_log.userId
```

```
Out[28]: [Row(userId='1138', firstname='Kelly', page='Home', level='paid', song=None),
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Everybody E
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Gears'),
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Use Somebod
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Love Of My
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Down In The
          Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Treat Her L
```

```
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="Everybody T
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Fourteen Wi
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Love On The
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Breakeven')
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Leaf House'
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='NAISEN KANS
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="You're In M
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Roll On Dow
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Plasticitie
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Secrets'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Hello'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='I Never Tol
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Love Break
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='One Touch O
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Undo'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Overdue (Bl
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Slave To Lo
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Stronger'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='All Of Us (
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Sehr kosmis
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='March Of Th
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Electricity
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Aces High')
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Bananeira')
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='The General
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='HÃ\x83Âľroe
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="Don't Stop
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="You're The
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Entering Wh
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Piccolo Ces
Row(userId='1138', firstname='Kelly', page='Help', level='paid', song=None),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Last Christ
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='You Shook M
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Going Stead
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='My Name Is'
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Undo'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Secrets'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Good Times
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Angelito'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Batdance (
Row(userId='1138', firstname='Kelly', page='Home', level='paid', song=None),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='DiÃ\x83Âąkd
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Whirring'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Potholderz
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Seaside'),
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Louder Than
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Just Like Y
Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="You're The
```

```
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Turn It Aga
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Everywhere
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song="Easy Skanki
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Roses'),
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Killing Me
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='The Razor (
 Row(userId='1138', firstname='Kelly', page='NextSong', level='paid', song='Idols and A
 Row(userId='1138', firstname='Kelly', page='Downgrade', level='paid', song=None),
 Row(userId='1138', firstname='Kelly', page='Submit Downgrade', level='paid', song=None
 Row(userId='1138', firstname='Kelly', page='Home', level='free', song=None),
 Row(userId='1138', firstname='Kelly', page='NextSong', level='free', song='Bones'),
 Row(userId='1138', firstname='Kelly', page='Home', level='free', song=None),
 Row(userId='1138', firstname='Kelly', page='NextSong', level='free', song='Grenouilles
```

```
In [29]: flag_downgrade_event = udf(lambda x: 1 if x == "Submit Downgrade" else 0, IntegerType()
```

```
In [30]: user_log_valid = user_log_valid.withColumn("downgraded", flag_downgrade_event("page"))
```

```
In [31]: user_log_valid.head()
```

```
Out[31]: Row(artist='Showaddywaddy', auth='Logged In', firstName='Kenneth', gender='M', itemInSe
```

```
In [32]: from pyspark.sql import Window
```

```
In [33]: windowval = Window.partitionBy("userId").orderBy(desc("ts")).rangeBetween(Window.unboun
```

```
In [34]: user_log_valid = user_log_valid.withColumn("phase", Fsum("downgraded").over(windowval))
```

```
In [35]: user_log_valid.select(["userId", "firstname", "ts", "page", "level", "phase"]).where(us
```

```
Out[35]: [Row(userId='1138', firstname='Kelly', ts=1513729066284, page='Home', level='paid', pha
 Row(userId='1138', firstname='Kelly', ts=1513729066284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513729313284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513729552284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513729783284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513730001284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513730263284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513730518284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513730768284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513731182284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513731435284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513731695284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513731857284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513732160284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513732302284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513732540284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513732770284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513732994284, page='NextSong', level='paid',
 Row(userId='1138', firstname='Kelly', ts=1513733223284, page='NextSong', level='paid',
```

```
Row(userId='1138', firstname='Kelly', ts=1513733456284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513733738284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513733941284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513734289284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513734598284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513734863284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513735174284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513735385284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513736040284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513736237284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513736446284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513736709284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513736915284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513737160284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513737460284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513737728284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513737936284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513738144284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513738197284, page='Help', level='paid', pha
Row(userId='1138', firstname='Kelly', ts=1513738432284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513763195284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513763453284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513763622284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513763890284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513764238284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513764462284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513764782284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513765075284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513765075284, page='Home', level='paid', pha
Row(userId='1138', firstname='Kelly', ts=1513765448284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513765655284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513765818284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513766091284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513766189284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513766385284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513766599284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513766838284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513767203284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513767413284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513767643284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513768012284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513768242284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513768452284, page='NextSong', level='paid',
Row(userId='1138', firstname='Kelly', ts=1513768453284, page='Downgrade', level='paid'
Row(userId='1138', firstname='Kelly', ts=1513768454284, page='Submit Downgrade', level
Row(userId='1138', firstname='Kelly', ts=1513768456284, page='Home', level='free', pha
Row(userId='1138', firstname='Kelly', ts=1513814880284, page='NextSong', level='free',
Row(userId='1138', firstname='Kelly', ts=1513821430284, page='Home', level='free', pha
```

```
          Row(userId='1138', firstname='Kelly', ts=1513833144284, page='NextSong', level='free',
```

```
In [37]: user_log.select(["userId"]).dropDuplicates().where(user_log.gender == 'F').count()
```

```
Out[37]: 462
```

```
In [38]: user_log.select(["artist"]).groupby(user_log.artist).count().sort(desc("count")).show()
```

```
+--------------------+-----+
|              artist|count|
+--------------------+-----+
|                null| 1653|
|            Coldplay|   83|
|       Kings Of Leon|   69|
|Florence + The Ma...|   52|
|            BjÃČÂűrk|   46|
|       Dwight Yoakam|   45|
|       Justin Bieber|   43|
|      The Black Keys|   40|
|         OneRepublic|   37|
|                Muse|   36|
|        Jack Johnson|   36|
|           Radiohead|   31|
|        Taylor Swift|   29|
|Barry Tuckwell/Ac...|   28|
|          Lily Allen|   28|
|               Train|   28|
|           Daft Punk|   27|
|           Metallica|   27|
|          Nickelback|   27|
|          Kanye West|   26|
+--------------------+-----+
only showing top 20 rows
```

```
In [47]: user_log.select(["page"]).groupby(user_log.page).count().collect()
```

```
Out[47]: [Row(page='Submit Downgrade', count=1),
          Row(page='Home', count=1126),
          Row(page='Downgrade', count=75),
          Row(page='Logout', count=100),
          Row(page='Save Settings', count=11),
          Row(page='About', count=43),
          Row(page='Settings', count=59),
          Row(page='Login', count=126),
          Row(page='NextSong', count=8347),
          Row(page='Help', count=58),
          Row(page='Upgrade', count=32),
```

```
          Row(page='Error', count=12),
          Row(page='Submit Upgrade', count=10)]

In [41]: user_log.select(["artist", "song"]).filter("artist = 'Coldplay'").count()

Out[41]: 83

In [61]: user_log.select(["page", "userId"]).filter("page = 'Home'").dropDuplicates().collect()

Out[61]: [Row(page='Home', userId='926'),
          Row(page='Home', userId='986'),
          Row(page='Home', userId='2402'),
          Row(page='Home', userId='2231'),
          Row(page='Home', userId='1701'),
          Row(page='Home', userId='1854'),
          Row(page='Home', userId='2354'),
          Row(page='Home', userId='2994'),
          Row(page='Home', userId='1708'),
          Row(page='Home', userId='1779'),
          Row(page='Home', userId='950'),
          Row(page='Home', userId='1280'),
          Row(page='Home', userId='2980'),
          Row(page='Home', userId='1312'),
          Row(page='Home', userId='2020'),
          Row(page='Home', userId='2694'),
          Row(page='Home', userId='577'),
          Row(page='Home', userId='2202'),
          Row(page='Home', userId='541'),
          Row(page='Home', userId='1944'),
          Row(page='Home', userId='2244'),
          Row(page='Home', userId='1174'),
          Row(page='Home', userId='2274'),
          Row(page='Home', userId='476'),
          Row(page='Home', userId='1199'),
          Row(page='Home', userId='298'),
          Row(page='Home', userId='471'),
          Row(page='Home', userId='2589'),
          Row(page='Home', userId='678'),
          Row(page='Home', userId='2125'),
          Row(page='Home', userId='712'),
          Row(page='Home', userId='1780'),
          Row(page='Home', userId='1968'),
          Row(page='Home', userId='2553'),
          Row(page='Home', userId='2767'),
          Row(page='Home', userId='937'),
          Row(page='Home', userId='1374'),
          Row(page='Home', userId='2239'),
          Row(page='Home', userId='699'),
          Row(page='Home', userId='2313'),
```

```
Row(page='Home', userId='486'),
Row(page='Home', userId='1926'),
Row(page='Home', userId='2469'),
Row(page='Home', userId='750'),
Row(page='Home', userId='2233'),
Row(page='Home', userId='1140'),
Row(page='Home', userId='1810'),
Row(page='Home', userId='961'),
Row(page='Home', userId='79'),
Row(page='Home', userId='1050'),
Row(page='Home', userId='249'),
Row(page='Home', userId='438'),
Row(page='Home', userId='1082'),
Row(page='Home', userId='1664'),
Row(page='Home', userId='196'),
Row(page='Home', userId='2132'),
Row(page='Home', userId='649'),
Row(page='Home', userId='2933'),
Row(page='Home', userId='2795'),
Row(page='Home', userId='393'),
Row(page='Home', userId='2268'),
Row(page='Home', userId='130'),
Row(page='Home', userId='1350'),
Row(page='Home', userId='685'),
Row(page='Home', userId='2816'),
Row(page='Home', userId='2615'),
Row(page='Home', userId='1074'),
Row(page='Home', userId='1949'),
Row(page='Home', userId='533'),
Row(page='Home', userId='1423'),
Row(page='Home', userId='1494'),
Row(page='Home', userId='1172'),
Row(page='Home', userId='805'),
Row(page='Home', userId='1393'),
Row(page='Home', userId='2696'),
Row(page='Home', userId='2049'),
Row(page='Home', userId='2636'),
Row(page='Home', userId='2533'),
Row(page='Home', userId='1964'),
Row(page='Home', userId='2528'),
Row(page='Home', userId='884'),
Row(page='Home', userId='1869'),
Row(page='Home', userId='2115'),
Row(page='Home', userId='1803'),
Row(page='Home', userId='243'),
Row(page='Home', userId='165'),
Row(page='Home', userId='2462'),
Row(page='Home', userId='1272'),
```

```
Row(page='Home', userId='1697'),
Row(page='Home', userId='1182'),
Row(page='Home', userId='1205'),
Row(page='Home', userId='2713'),
Row(page='Home', userId='1847'),
Row(page='Home', userId='136'),
Row(page='Home', userId='1595'),
Row(page='Home', userId='1543'),
Row(page='Home', userId='2311'),
Row(page='Home', userId='1887'),
Row(page='Home', userId='2853'),
Row(page='Home', userId='625'),
Row(page='Home', userId='2880'),
Row(page='Home', userId='2964'),
Row(page='Home', userId='450'),
Row(page='Home', userId='1569'),
Row(page='Home', userId='1399'),
Row(page='Home', userId='644'),
Row(page='Home', userId='1503'),
Row(page='Home', userId='1290'),
Row(page='Home', userId='410'),
Row(page='Home', userId='1642'),
Row(page='Home', userId='2708'),
Row(page='Home', userId='2551'),
Row(page='Home', userId='1522'),
Row(page='Home', userId='1790'),
Row(page='Home', userId='1961'),
Row(page='Home', userId='2721'),
Row(page='Home', userId='275'),
Row(page='Home', userId='60'),
Row(page='Home', userId='2367'),
Row(page='Home', userId='2897'),
Row(page='Home', userId='1580'),
Row(page='Home', userId='517'),
Row(page='Home', userId='2217'),
Row(page='Home', userId='989'),
Row(page='Home', userId=''),
Row(page='Home', userId='1625'),
Row(page='Home', userId='1398'),
Row(page='Home', userId='2634'),
Row(page='Home', userId='2111'),
Row(page='Home', userId='2213'),
Row(page='Home', userId='1478'),
Row(page='Home', userId='1380'),
Row(page='Home', userId='361'),
Row(page='Home', userId='1911'),
Row(page='Home', userId='1975'),
Row(page='Home', userId='912'),
```

```
Row(page='Home', userId='2251'),
Row(page='Home', userId='1990'),
Row(page='Home', userId='88'),
Row(page='Home', userId='2096'),
Row(page='Home', userId='691'),
Row(page='Home', userId='1691'),
Row(page='Home', userId='1551'),
Row(page='Home', userId='1604'),
Row(page='Home', userId='11'),
Row(page='Home', userId='970'),
Row(page='Home', userId='2902'),
Row(page='Home', userId='1177'),
Row(page='Home', userId='1412'),
Row(page='Home', userId='321'),
Row(page='Home', userId='2612'),
Row(page='Home', userId='1319'),
Row(page='Home', userId='647'),
Row(page='Home', userId='211'),
Row(page='Home', userId='1107'),
Row(page='Home', userId='2945'),
Row(page='Home', userId='2105'),
Row(page='Home', userId='1347'),
Row(page='Home', userId='187'),
Row(page='Home', userId='2775'),
Row(page='Home', userId='429'),
Row(page='Home', userId='979'),
Row(page='Home', userId='2186'),
Row(page='Home', userId='1077'),
Row(page='Home', userId='1155'),
Row(page='Home', userId='1430'),
Row(page='Home', userId='2567'),
Row(page='Home', userId='597'),
Row(page='Home', userId='2200'),
Row(page='Home', userId='888'),
Row(page='Home', userId='2455'),
Row(page='Home', userId='374'),
Row(page='Home', userId='1943'),
Row(page='Home', userId='1514'),
Row(page='Home', userId='2990'),
Row(page='Home', userId='1568'),
Row(page='Home', userId='58'),
Row(page='Home', userId='266'),
Row(page='Home', userId='354'),
Row(page='Home', userId='1141'),
Row(page='Home', userId='822'),
Row(page='Home', userId='2079'),
Row(page='Home', userId='247'),
Row(page='Home', userId='2018'),
```

```
Row(page='Home', userId='1842'),
Row(page='Home', userId='1828'),
Row(page='Home', userId='900'),
Row(page='Home', userId='297'),
Row(page='Home', userId='308'),
Row(page='Home', userId='1219'),
Row(page='Home', userId='2187'),
Row(page='Home', userId='1881'),
Row(page='Home', userId='55'),
Row(page='Home', userId='1019'),
Row(page='Home', userId='105'),
Row(page='Home', userId='2884'),
Row(page='Home', userId='590'),
Row(page='Home', userId='2204'),
Row(page='Home', userId='942'),
Row(page='Home', userId='343'),
Row(page='Home', userId='1046'),
Row(page='Home', userId='896'),
Row(page='Home', userId='1540'),
Row(page='Home', userId='1647'),
Row(page='Home', userId='807'),
Row(page='Home', userId='1383'),
Row(page='Home', userId='1079'),
Row(page='Home', userId='2417'),
Row(page='Home', userId='2593'),
Row(page='Home', userId='1153'),
Row(page='Home', userId='316'),
Row(page='Home', userId='726'),
Row(page='Home', userId='1865'),
Row(page='Home', userId='226'),
Row(page='Home', userId='1806'),
Row(page='Home', userId='578'),
Row(page='Home', userId='1906'),
Row(page='Home', userId='2734'),
Row(page='Home', userId='2599'),
Row(page='Home', userId='967'),
Row(page='Home', userId='2786'),
Row(page='Home', userId='2234'),
Row(page='Home', userId='800'),
Row(page='Home', userId='43'),
Row(page='Home', userId='2117'),
Row(page='Home', userId='526'),
Row(page='Home', userId='568'),
Row(page='Home', userId='1860'),
Row(page='Home', userId='252'),
Row(page='Home', userId='282'),
Row(page='Home', userId='855'),
Row(page='Home', userId='2703'),
```

```
Row(page='Home', userId='2073'),
Row(page='Home', userId='630'),
Row(page='Home', userId='1035'),
Row(page='Home', userId='1820'),
Row(page='Home', userId='540'),
Row(page='Home', userId='109'),
Row(page='Home', userId='1588'),
Row(page='Home', userId='2290'),
Row(page='Home', userId='2376'),
Row(page='Home', userId='2481'),
Row(page='Home', userId='1724'),
Row(page='Home', userId='2489'),
Row(page='Home', userId='2297'),
Row(page='Home', userId='534'),
Row(page='Home', userId='2552'),
Row(page='Home', userId='2459'),
Row(page='Home', userId='779'),
Row(page='Home', userId='441'),
Row(page='Home', userId='1608'),
Row(page='Home', userId='1586'),
Row(page='Home', userId='465'),
Row(page='Home', userId='1916'),
Row(page='Home', userId='338'),
Row(page='Home', userId='974'),
Row(page='Home', userId='1583'),
Row(page='Home', userId='1065'),
Row(page='Home', userId='2411'),
Row(page='Home', userId='656'),
Row(page='Home', userId='2388'),
Row(page='Home', userId='2777'),
Row(page='Home', userId='2759'),
Row(page='Home', userId='1497'),
Row(page='Home', userId='367'),
Row(page='Home', userId='1413'),
Row(page='Home', userId='1891'),
Row(page='Home', userId='981'),
Row(page='Home', userId='2946'),
Row(page='Home', userId='1138'),
Row(page='Home', userId='2949'),
Row(page='Home', userId='2275'),
Row(page='Home', userId='1629'),
Row(page='Home', userId='1986'),
Row(page='Home', userId='80'),
Row(page='Home', userId='1645'),
Row(page='Home', userId='1429'),
Row(page='Home', userId='2176'),
Row(page='Home', userId='1952'),
Row(page='Home', userId='138'),
```

```
Row(page='Home', userId='2305'),
Row(page='Home', userId='1950'),
Row(page='Home', userId='1377'),
Row(page='Home', userId='2147'),
Row(page='Home', userId='1109'),
Row(page='Home', userId='877'),
Row(page='Home', userId='2950'),
Row(page='Home', userId='2219'),
Row(page='Home', userId='2294'),
Row(page='Home', userId='62'),
Row(page='Home', userId='2359'),
Row(page='Home', userId='1661'),
Row(page='Home', userId='1243'),
Row(page='Home', userId='2535'),
Row(page='Home', userId='893'),
Row(page='Home', userId='785'),
Row(page='Home', userId='2715'),
Row(page='Home', userId='1281'),
Row(page='Home', userId='1915'),
Row(page='Home', userId='47'),
Row(page='Home', userId='1309'),
Row(page='Home', userId='1840'),
Row(page='Home', userId='1062'),
Row(page='Home', userId='2732'),
Row(page='Home', userId='1829'),
Row(page='Home', userId='553'),
Row(page='Home', userId='2743'),
Row(page='Home', userId='229'),
Row(page='Home', userId='1786'),
Row(page='Home', userId='803'),
Row(page='Home', userId='1903'),
Row(page='Home', userId='2675'),
Row(page='Home', userId='158'),
Row(page='Home', userId='1991'),
Row(page='Home', userId='724'),
Row(page='Home', userId='2387'),
Row(page='Home', userId='40'),
Row(page='Home', userId='392'),
Row(page='Home', userId='150'),
Row(page='Home', userId='2047'),
Row(page='Home', userId='2600'),
Row(page='Home', userId='387'),
Row(page='Home', userId='422'),
Row(page='Home', userId='2479'),
Row(page='Home', userId='254'),
Row(page='Home', userId='1845'),
Row(page='Home', userId='2603'),
Row(page='Home', userId='1043'),
```

```
Row(page='Home', userId='2181'),
Row(page='Home', userId='2168'),
Row(page='Home', userId='2033'),
Row(page='Home', userId='2909'),
Row(page='Home', userId='846'),
Row(page='Home', userId='2349'),
Row(page='Home', userId='2801'),
Row(page='Home', userId='1584'),
Row(page='Home', userId='442'),
Row(page='Home', userId='2967'),
Row(page='Home', userId='1295'),
Row(page='Home', userId='1564'),
Row(page='Home', userId='2247'),
Row(page='Home', userId='59'),
Row(page='Home', userId='1611'),
Row(page='Home', userId='2088'),
Row(page='Home', userId='1471'),
Row(page='Home', userId='1777'),
Row(page='Home', userId='2594'),
Row(page='Home', userId='622'),
Row(page='Home', userId='2756'),
Row(page='Home', userId='364'),
Row(page='Home', userId='2416'),
Row(page='Home', userId='523'),
Row(page='Home', userId='1809'),
Row(page='Home', userId='379'),
Row(page='Home', userId='1392'),
Row(page='Home', userId='186'),
Row(page='Home', userId='1426'),
Row(page='Home', userId='571'),
Row(page='Home', userId='2664'),
Row(page='Home', userId='2737'),
Row(page='Home', userId='1747'),
Row(page='Home', userId='732'),
Row(page='Home', userId='2109'),
Row(page='Home', userId='32'),
Row(page='Home', userId='804'),
Row(page='Home', userId='2372'),
Row(page='Home', userId='141'),
Row(page='Home', userId='391'),
Row(page='Home', userId='1233'),
Row(page='Home', userId='362'),
Row(page='Home', userId='314'),
Row(page='Home', userId='2113'),
Row(page='Home', userId='117'),
Row(page='Home', userId='902'),
Row(page='Home', userId='2975'),
Row(page='Home', userId='1561'),
```

```
Row(page='Home', userId='2940'),
Row(page='Home', userId='1959'),
Row(page='Home', userId='2162'),
Row(page='Home', userId='52'),
Row(page='Home', userId='2632'),
Row(page='Home', userId='2404'),
Row(page='Home', userId='1918'),
Row(page='Home', userId='446'),
Row(page='Home', userId='1598'),
Row(page='Home', userId='2391'),
Row(page='Home', userId='255'),
Row(page='Home', userId='2053'),
Row(page='Home', userId='2085'),
Row(page='Home', userId='700'),
Row(page='Home', userId='420'),
Row(page='Home', userId='825'),
Row(page='Home', userId='2882'),
Row(page='Home', userId='2731'),
Row(page='Home', userId='1020'),
Row(page='Home', userId='2346'),
Row(page='Home', userId='1394'),
Row(page='Home', userId='2906'),
Row(page='Home', userId='1048'),
Row(page='Home', userId='949'),
Row(page='Home', userId='1885'),
Row(page='Home', userId='1117'),
Row(page='Home', userId='323'),
Row(page='Home', userId='2748'),
Row(page='Home', userId='302'),
Row(page='Home', userId='1158'),
Row(page='Home', userId='862'),
Row(page='Home', userId='1349'),
Row(page='Home', userId='880'),
Row(page='Home', userId='623'),
Row(page='Home', userId='2412'),
Row(page='Home', userId='1367'),
Row(page='Home', userId='2065'),
Row(page='Home', userId='1267'),
Row(page='Home', userId='2927'),
Row(page='Home', userId='451'),
Row(page='Home', userId='2038'),
Row(page='Home', userId='1025'),
Row(page='Home', userId='1017'),
Row(page='Home', userId='2852'),
Row(page='Home', userId='1725'),
Row(page='Home', userId='2017'),
Row(page='Home', userId='1322'),
Row(page='Home', userId='64'),
```

```
Row(page='Home', userId='1303'),
Row(page='Home', userId='2764'),
Row(page='Home', userId='1688'),
Row(page='Home', userId='1232'),
Row(page='Home', userId='790'),
Row(page='Home', userId='2426'),
Row(page='Home', userId='2617'),
Row(page='Home', userId='2922'),
Row(page='Home', userId='1040'),
Row(page='Home', userId='1103'),
Row(page='Home', userId='1541'),
Row(page='Home', userId='1368'),
Row(page='Home', userId='707'),
Row(page='Home', userId='35'),
Row(page='Home', userId='2028'),
Row(page='Home', userId='462'),
Row(page='Home', userId='913'),
Row(page='Home', userId='773'),
Row(page='Home', userId='978'),
Row(page='Home', userId='2549'),
Row(page='Home', userId='1165'),
Row(page='Home', userId='174'),
Row(page='Home', userId='100'),
Row(page='Home', userId='2842'),
Row(page='Home', userId='892'),
Row(page='Home', userId='1980'),
Row(page='Home', userId='2515'),
Row(page='Home', userId='2074'),
Row(page='Home', userId='1795'),
Row(page='Home', userId='2160'),
Row(page='Home', userId='2892'),
Row(page='Home', userId='1042'),
Row(page='Home', userId='2628'),
Row(page='Home', userId='613'),
Row(page='Home', userId='648'),
Row(page='Home', userId='480'),
Row(page='Home', userId='2815'),
Row(page='Home', userId='2885'),
Row(page='Home', userId='1678'),
Row(page='Home', userId='2958'),
Row(page='Home', userId='2988'),
Row(page='Home', userId='1680'),
Row(page='Home', userId='1422'),
Row(page='Home', userId='1967'),
Row(page='Home', userId='1573'),
Row(page='Home', userId='2991'),
Row(page='Home', userId='1818'),
Row(page='Home', userId='1819'),
```

```
Row(page='Home', userId='2273'),
Row(page='Home', userId='1208'),
Row(page='Home', userId='1461'),
Row(page='Home', userId='1037'),
Row(page='Home', userId='2081'),
Row(page='Home', userId='1969'),
Row(page='Home', userId='1684'),
Row(page='Home', userId='2383'),
Row(page='Home', userId='2293'),
Row(page='Home', userId='639'),
Row(page='Home', userId='377'),
Row(page='Home', userId='2122'),
Row(page='Home', userId='2304'),
Row(page='Home', userId='2813'),
Row(page='Home', userId='680'),
Row(page='Home', userId='2873'),
Row(page='Home', userId='2996'),
Row(page='Home', userId='2574'),
Row(page='Home', userId='527'),
Row(page='Home', userId='2022'),
Row(page='Home', userId='635'),
Row(page='Home', userId='869'),
Row(page='Home', userId='1152'),
Row(page='Home', userId='512'),
Row(page='Home', userId='1452'),
Row(page='Home', userId='957'),
Row(page='Home', userId='1852'),
Row(page='Home', userId='1523'),
Row(page='Home', userId='2259'),
Row(page='Home', userId='1003'),
Row(page='Home', userId='2243'),
Row(page='Home', userId='305'),
Row(page='Home', userId='2373'),
Row(page='Home', userId='1373'),
Row(page='Home', userId='1929'),
Row(page='Home', userId='2829'),
Row(page='Home', userId='1970'),
Row(page='Home', userId='2800'),
Row(page='Home', userId='331'),
Row(page='Home', userId='1741'),
Row(page='Home', userId='1098'),
Row(page='Home', userId='1668'),
Row(page='Home', userId='1346'),
Row(page='Home', userId='1715'),
Row(page='Home', userId='2095'),
Row(page='Home', userId='2912'),
Row(page='Home', userId='2830'),
Row(page='Home', userId='936'),
```

```
Row(page='Home', userId='2143'),
Row(page='Home', userId='856'),
Row(page='Home', userId='2997'),
Row(page='Home', userId='748'),
Row(page='Home', userId='1746'),
Row(page='Home', userId='598'),
Row(page='Home', userId='2319'),
Row(page='Home', userId='1998'),
Row(page='Home', userId='814'),
Row(page='Home', userId='1776'),
Row(page='Home', userId='2333'),
Row(page='Home', userId='2164'),
Row(page='Home', userId='1330'),
Row(page='Home', userId='2760'),
Row(page='Home', userId='1278'),
Row(page='Home', userId='1358'),
Row(page='Home', userId='2749'),
Row(page='Home', userId='2425'),
Row(page='Home', userId='1176'),
Row(page='Home', userId='2993'),
Row(page='Home', userId='628'),
Row(page='Home', userId='1083'),
Row(page='Home', userId='99'),
Row(page='Home', userId='2921'),
Row(page='Home', userId='1520'),
Row(page='Home', userId='821'),
Row(page='Home', userId='574'),
Row(page='Home', userId='505'),
Row(page='Home', userId='2004'),
Row(page='Home', userId='2845'),
Row(page='Home', userId='702'),
Row(page='Home', userId='2791'),
Row(page='Home', userId='1061'),
Row(page='Home', userId='28'),
Row(page='Home', userId='1644'),
Row(page='Home', userId='757'),
Row(page='Home', userId='2570'),
Row(page='Home', userId='2051'),
Row(page='Home', userId='1855'),
Row(page='Home', userId='1490'),
Row(page='Home', userId='2665'),
Row(page='Home', userId='910'),
Row(page='Home', userId='2531'),
Row(page='Home', userId='1163'),
Row(page='Home', userId='325'),
Row(page='Home', userId='746'),
Row(page='Home', userId='2539'),
Row(page='Home', userId='159'),
```

```
 Row(page='Home', userId='2692'),
 Row(page='Home', userId='2799'),
 Row(page='Home', userId='2418'),
 Row(page='Home', userId='1558'),
 Row(page='Home', userId='2944'),
 Row(page='Home', userId='2089'),
 Row(page='Home', userId='210'),
 Row(page='Home', userId='1686'),
 Row(page='Home', userId='2652'),
 Row(page='Home', userId='496'),
 Row(page='Home', userId='2685'),
 Row(page='Home', userId='2035'),
 Row(page='Home', userId='1396'),
 Row(page='Home', userId='2021'),
 Row(page='Home', userId='2048'),
 Row(page='Home', userId='1567'),
 Row(page='Home', userId='1264'),
 Row(page='Home', userId='1362'),
 Row(page='Home', userId='395'),
 Row(page='Home', userId='1238'),
 Row(page='Home', userId='1955'),
 Row(page='Home', userId='2190'),
 Row(page='Home', userId='1331'),
 Row(page='Home', userId='1920'),
 Row(page='Home', userId='1442'),
 Row(page='Home', userId='1069'),
 Row(page='Home', userId='2076'),
 Row(page='Home', userId='2322'),
 Row(page='Home', userId='2963'),
 Row(page='Home', userId='761'),
 Row(page='Home', userId='1388'),
 Row(page='Home', userId='1132'),
 Row(page='Home', userId='830'),
 Row(page='Home', userId='61')]
```

In [74]:
```
li=['Home','NextSong']
user_log.select(["page", "userId"]).filter(user_log.page.isin(li)).dropDuplicates().cou
```

Out[74]: 1447

In [ ]: