What happens in Vegas, Stay in Venmo

Team: Fantastic Five (Vincy Hu, Nandini Basu, Frank He, Aijie Li, Allie Tsuji)

I. Text Analysis

Q0: Your first task is to open your Venmo app, find 10 words that are not already in the dictionary, and add them to it. Make sure you don't add to the dictionary a duplicate word by hitting Control+F before adding your word.

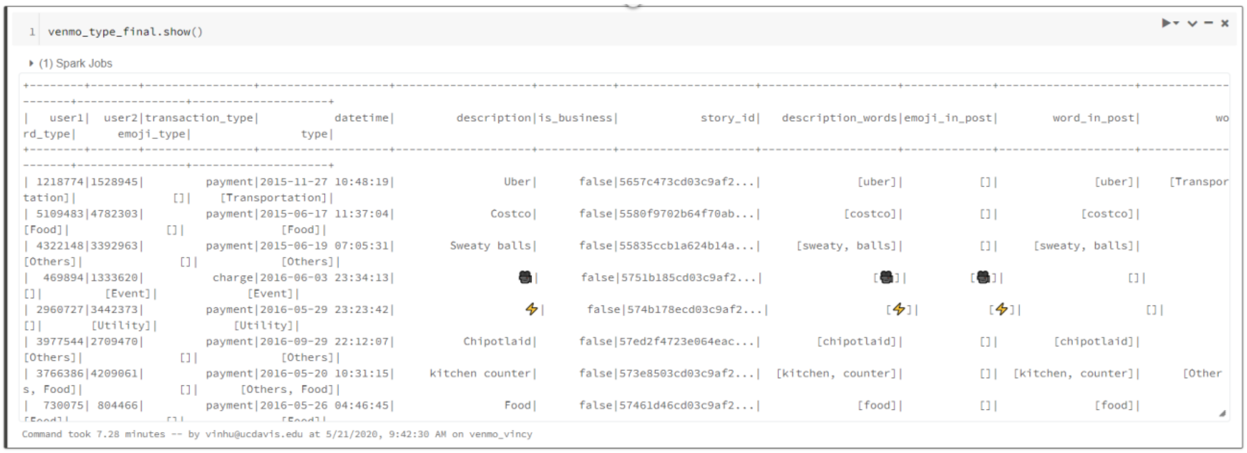
The word our group added in the document are:
scream [people], KFC [food], tmobile [utility], hmart [food], watch [cash], hat [cash], supreme [cash], spinach [food], hall [food], spaghetti [food]

Q1: Use the text dictionary and the emoji dictionary to classify Venmo's transactions in your sample dataset.

After importing the word and emoji dictionary, we tokenized the description value to lists of word/emoji elements. Then, we separate word with emoji to have column'word_in_post' and column 'emoji_in_post'. We compare word with word dictionary, emoji with emoji dictionary to get the total categories appeared in the description in each record.

Assumption here: for each transaction, there can be multiple categories, but each category only counts for one time even if multiple words for the same category appear more than once.

Please refer to the notebook for detail result:



```
1  venmo_type_final.show()
```

▸ (1) Spark Jobs

```
+--------+-------+---------------+-------------------+--------------------+----------+--------------------+-------------------+------------+--------------------+------------
-------+---------------+-------------------+
|   user1|  user2|transaction_type|           datetime|         description|is_business|            story_id|   description_words|emoji_in_post|        word_in_post|          wo
rd_type|      emoji_type|               type|
+--------+-------+---------------+-------------------+--------------------+----------+--------------------+-------------------+------------+--------------------+------------
-------+---------------+-------------------+
| 1218774|1528945|        payment|2015-11-27 10:48:19|                Uber|     false|5657c473cd03c9af2...|             [uber]|          []|              [uber]|   [Transpor
tation]|             []|    [Transportation]|
| 5109483|4782303|        payment|2015-06-17 11:37:04|              Costco|     false|5580f9702b64f70ab...|           [costco]|          []|            [costco]|
[Food]|             []|             [Food]|
| 4322148|3392963|        payment|2015-06-19 07:05:31|        Sweaty balls|     false|55835ccb1a624b14a...|    [sweaty, balls]|          []|     [sweaty, balls]|
[Others]|             []|           [Others]|
|  469894|1333620|         charge|2016-06-03 23:34:13|                  📷|     false|5751b185cd03c9af2...|               [📷]|        [📷]|                  []|
[]|        [Event]|            [Event]|
| 2960727|3442373|        payment|2016-05-29 23:23:42|                   ⚡|     false|574b178ecd03c9af2...|                [⚡]|         [⚡]|                  []|
[]|      [Utility]|          [Utility]|
| 3977544|2709470|        payment|2016-09-29 22:12:07|           Chipotlaid|     false|57ed2f4723e064eac...|       [chipotlaid]|          []|        [chipotlaid]|
[Others]|             []|           [Others]|
| 3766386|4209061|        payment|2016-05-20 10:31:15|     kitchen counter|     false|573e8503cd03c9af2...| [kitchen, counter]|          []|  [kitchen, counter]|      [Other
s, Food]|             []|     [Others, Food]|
|  730075| 804466|        payment|2016-05-26 04:46:45|                Food|     false|57461d46cd03c9af2...|             [food]|          []|              [food]|
[Food]|             []|             [Food]|
```

Command took 7.28 minutes -- by vinhu@ucdavis.edu at 5/21/2020, 9:42:30 AM on venmo_vincy

Q2: What is the percent of emoji only transactions? Which are the top 5 most popular emoji? Which are the top three most popular emoji categories?

To detect emoji only transaction, we write functions to return whether there's word and whether there is emoji in the description. When there's no word and there is emoji(s), the description will be counted.

Overall, we found around 30% of transactions only include emoji in the description. (That's a lot! 😬)

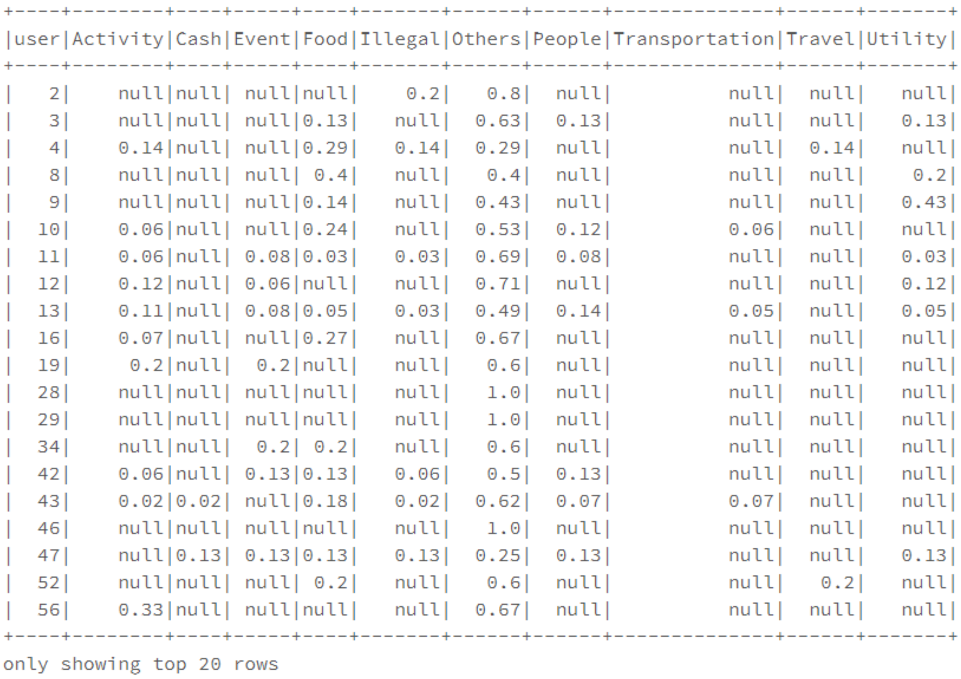
```
+----------+-------------+------------------+
|emoji_only|per_per_group|         per_total|
+----------+-------------+------------------+
|     False|      4924907| 69.23677977803605|
|      True|      2188230|30.763220221963948|
+----------+-------------+------------------+
```

The top five used emojis are:

```
+-----+------+
|emoji| count|
+-----+------+
|   🍕|215039|
|   🍺|145233|
|   💸|124727|
|   🍷|111157|
|   🎉| 94327|
+-----+------+
only showing top 5 rows
```

Q3: For each user, create a variable to indicate their spending behavior profile. For example, if a user has made 10 transactions, where 5 of them are food and the other 5 are activity, then the user's spending profile will be 50% food and 50% activity.

We noticed that each record is in either type of transaction: {payment, charge}. Since we want to depict users' spending profiles, we only take into account the 'payment' transaction. Thus we take user1 in any transaction with type ='payment' and user2 in any transaction with type ='charge'.

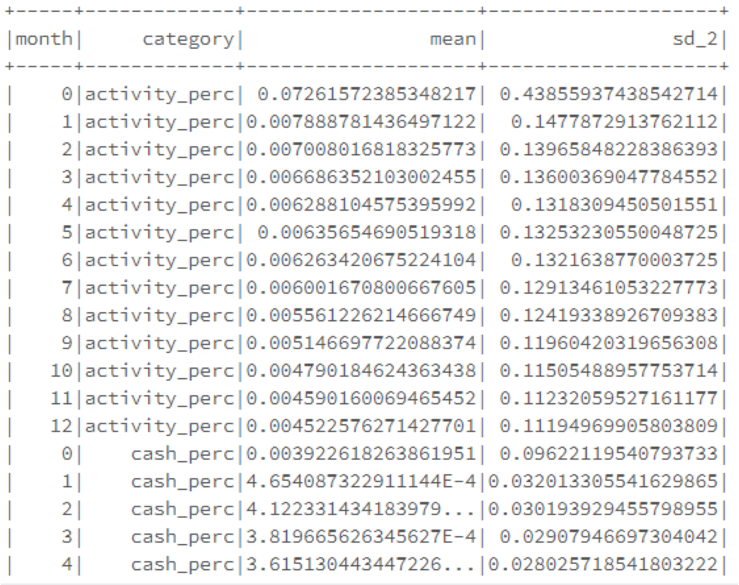


```
+----+--------+----+-----+----+-------+------+------+--------------+------+-------+
|user|Activity|Cash|Event|Food|Illegal|Others|People|Transportation|Travel|Utility|
+----+--------+----+-----+----+-------+------+------+--------------+------+-------+
|   2|    null|null| null|null|    0.2|   0.8|  null|          null|  null|   null|
|   3|    null|null| null|0.13|   null|  0.63|  0.13|          null|  null|   0.13|
|   4|    0.14|null| null|0.29|   0.14|  0.29|  null|          null|  0.14|   null|
|   8|    null|null| null| 0.4|   null|   0.4|  null|          null|  null|    0.2|
|   9|    null|null| null|0.14|   null|  0.43|  null|          null|  null|   0.43|
|  10|    0.06|null| null|0.24|   null|  0.53|  0.12|          0.06|  null|   null|
|  11|    0.06|null| 0.08|0.03|   0.03|  0.69|  0.08|          null|  null|   0.03|
|  12|    0.12|null| 0.06|null|   null|  0.71|  null|          null|  null|   0.12|
|  13|    0.11|null| 0.08|0.05|   0.03|  0.49|  0.14|          0.05|  null|   0.05|
|  16|    0.07|null| null|0.27|   null|  0.67|  null|          null|  null|   null|
|  19|     0.2|null|  0.2|null|   null|   0.6|  null|          null|  null|   null|
|  28|    null|null| null|null|   null|   1.0|  null|          null|  null|   null|
|  29|    null|null| null|null|   null|   1.0|  null|          null|  null|   null|
|  34|    null|null|  0.2| 0.2|   null|   0.6|  null|          null|  null|   null|
|  42|    0.06|null| 0.13|0.13|   0.06|   0.5|  0.13|          null|  null|   null|
|  43|    0.02|0.02| null|0.18|   0.02|  0.62|  0.07|          0.07|  null|   null|
|  46|    null|null| null|null|   null|   1.0|  null|          null|  null|   null|
|  47|    null|0.13| 0.13|0.13|   0.13|  0.25|  0.13|          null|  null|   0.13|
|  52|    null|null| null| 0.2|   null|   0.6|  null|          null|   0.2|   null|
|  56|    0.33|null| null|null|   null|  0.67|  null|          null|  null|   null|
+----+--------+----+-----+----+-------+------+------+--------------+------+-------+
only showing top 20 rows
```

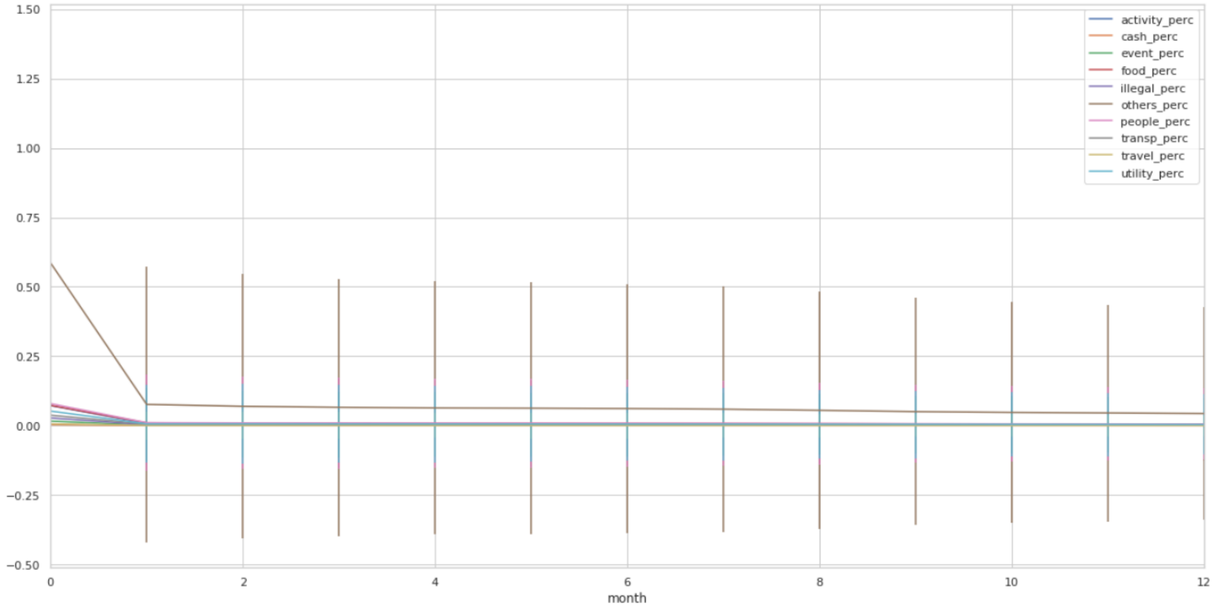
Please refer to the notebook for a detailed result.

Q4: In the previous question, you got a static spending profile. However, life and social networks are evolving over time. Therefore, let's explore how a user's spending profile is evolving over her lifetime in Venmo. First of all, you need to analyze a user's transactions in monthly intervals, starting from 0 (indicating their first transaction only) up to 12.

The following two graphs show the final result for calculating the mean and standard deviation for spending profiles of each category for each month among all users.



```
+-----+-------------+--------------------+--------------------+
|month|     category|                mean|                sd_2|
+-----+-------------+--------------------+--------------------+
|    0|activity_perc| 0.07261572385348217| 0.43855937438542714|
|    1|activity_perc|0.007888781436497122|  0.1477872913762112|
|    2|activity_perc|0.007008016818325773| 0.13965848228386393|
|    3|activity_perc|0.006686352103002455| 0.13600369047784552|
|    4|activity_perc|0.006288104575395992|  0.1318309450501551|
|    5|activity_perc| 0.00635654690519318| 0.13253230550048725|
|    6|activity_perc|0.006263420675224104|  0.1321638770003725|
|    7|activity_perc|0.006001670800667605| 0.12913461053227773|
|    8|activity_perc|0.005561226214666749| 0.12419338926709383|
|    9|activity_perc|0.005146697722088374| 0.11960420319656308|
|   10|activity_perc|0.004790184624363438| 0.11505488957753714|
|   11|activity_perc|0.004590160069465452| 0.11232059527161177|
|   12|activity_perc|0.004522576271427701| 0.11194969905803809|
|    0|    cash_perc|0.003922618263861951| 0.09622119540793733|
|    1|    cash_perc|4.654087322911144E-4|0.032013305541629865|
|    2|    cash_perc|4.12233143418397 9...|0.030193929455798955|
|    3|    cash_perc|3.819665626345627E-4| 0.02907946697304042|
|    4|    cash_perc|3.61513044344722 6...|0.028025718541803222|
```



According to the second graph, we noticed that most of the descriptions include words that cannot be identified by the current dictionary on our hand. In the time period 0, since it stands for every users' first transaction, the sum of categories should roughly sum to 1 cause there will be no user have 0 transaction in time 0. And almost 60% of transactions are identified as category 'other'.

Along 12-month lifetime, the mean of category 'other' slightly decreases and tends to stabilize after month 8, while keeping much the same for other categories. And the standard deviation becomes less. We assume this may be caused by some user churn, so mean declines and the left users' spending behavior become stable in the later period thus standard deviation declines too.

II. Social Network Analytics

Let's now look at a user's social network.

Q5: Write a script to find a user's friends and friends of friends (Friend definition: A user's friend is someone who has transacted with the user, either sending money to the user or receiving money from the user). Describe your algorithm and calculate its computational complexity. Can you do it better?

The assumption for this problem is that the two-side of the transaction are equally important, thus we union user1-user2 with user2-user1 to ensure not ignoring the user that only occurs in column 'user2'.

To solve this question, we tried different JOIN method to explore a relatively more efficient solution:

- First method:

  We tried to write SQL with spark API, and join the connection edge table('*<friending>*') with itself to find friend of each user and friend of a friend.

  By inspecting '.explain()' , we learned that the process has 14 steps for Spark to execute.
- Second method:

  Then, we tried Spark-specific SQL to join *<friending>* with itself. The physical plan shows that the steps are cut to 9. Much better!
- Third method:

  We learned from the textbook that the spark has communication strategy to deal with 'big data': Spark approaches cluster communication in two different ways during joins. It either incurs a *shuffle* join, which results in an all-to-all communication or a ***broadcast*** join.

  When we join a big table to another big table, it ends up with a shuffle join. In a shuffle join, every node talks to every other node and they share data according to which node has a certain key or set of keys (on which you are joining). These joins are expensive because the network can become congested with traffic. Thus, to make the join more efficient, we can approach the broadcast join.

By using the broadcast join, we finally land on 8 steps to do the join.

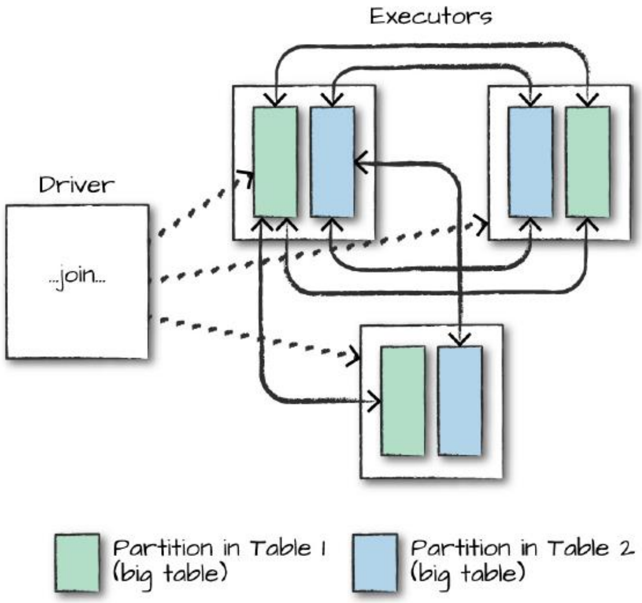

Executors
Driver
...join...
Partition in Table 1 (big table)
Partition in Table 2 (big table)

Figure 8-1. Joining two big tables

(Source: Chambers B. & Zaharia M. (2018) Spark, The Definitive Guide, Big Data Processing Made Simple. O'Reilly Media, Inc.)

Q6: Now, that you have the list of each user's friends and friends of friends, you are in position to calculate many social network variables. Use the dynamic analysis from before, and calculate the following social network metrics across a user's lifetime in Venmo (from 0 up to 12 months).

i) The number of friends and the number of friends of friends.

By friendship edge table from Question 5, we count distinct direct friends of each user and friend-of-friend for each month.

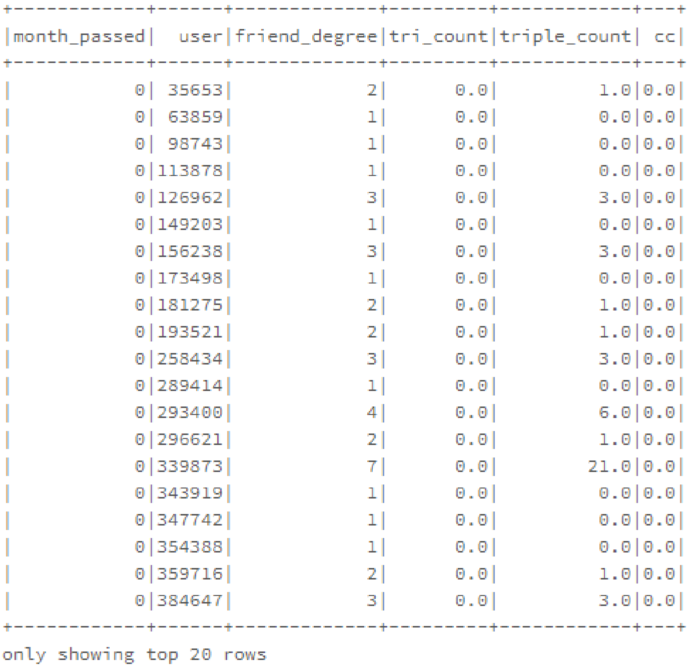
ii) The clustering coefficient of a user's network

Recall the definition of clustering coefficient:

CC(A) = P(two randomly selected friends of A are also friends)

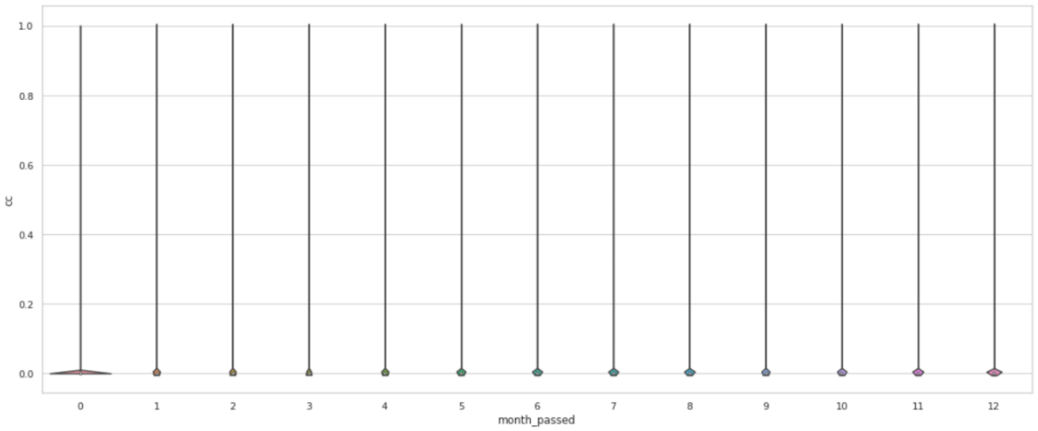
$$= \frac{\#\, of\ closed\ triangle\ of\ A(i.e.,where\ A\prime s\ two\ friends\ are\ also\ friends)}{\#\, of\ triples\ of\ A\ (any\ 2\ friends\ of\ A\ and\ A\ can\ combine\ a\ triplet)}$$

By joining the friending edge table with itself for 3 times, we can find each user's all closed triangle and divide number of triangles by triplets. Knowing each user's direct degree, the triplets number equals to combination Combination(degree of A, 2).



```
+------------+------+-------------+---------+------------+---+
|month_passed|  user|friend_degree|tri_count|triple_count| cc|
+------------+------+-------------+---------+------------+---+
|           0| 35653|            2|      0.0|         1.0|0.0|
|           0| 63859|            1|      0.0|         0.0|0.0|
|           0| 98743|            1|      0.0|         0.0|0.0|
|           0|113878|            1|      0.0|         0.0|0.0|
|           0|126962|            3|      0.0|         3.0|0.0|
|           0|149203|            1|      0.0|         0.0|0.0|
|           0|156238|            3|      0.0|         3.0|0.0|
|           0|173498|            1|      0.0|         0.0|0.0|
|           0|181275|            2|      0.0|         1.0|0.0|
|           0|193521|            2|      0.0|         1.0|0.0|
|           0|258434|            3|      0.0|         3.0|0.0|
|           0|289414|            1|      0.0|         0.0|0.0|
|           0|293400|            4|      0.0|         6.0|0.0|
|           0|296621|            2|      0.0|         1.0|0.0|
|           0|339873|            7|      0.0|        21.0|0.0|
|           0|343919|            1|      0.0|         0.0|0.0|
|           0|347742|            1|      0.0|         0.0|0.0|
|           0|354388|            1|      0.0|         0.0|0.0|
|           0|359716|            2|      0.0|         1.0|0.0|
|           0|384647|            3|      0.0|         3.0|0.0|
+------------+------+-------------+---------+------------+---+
only showing top 20 rows
```

We also noticed that across all the user, almost all the clustering coefficient are near zero, which mean that the network is relatively loose in a way that seldom do user's friend connect with each other.



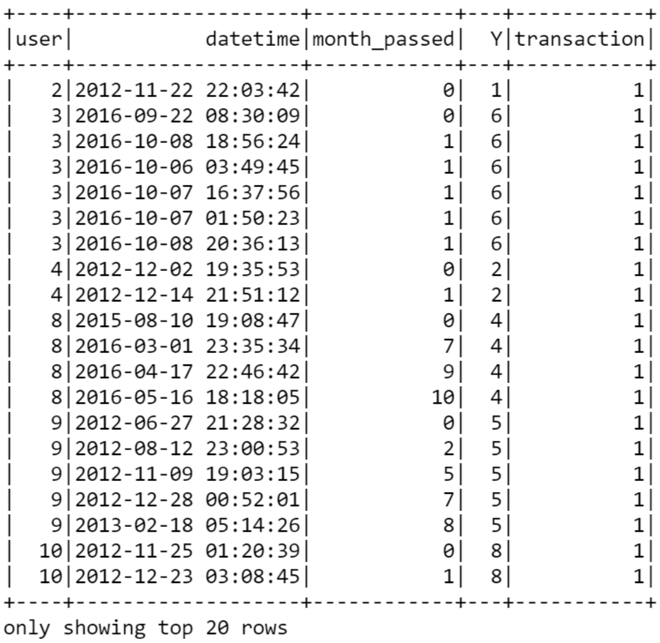
iii) Calculate the page rank of each user

Please refer to the notebook for detail.

III. Predictive Analytics with MLlib

Q7: First, create your dependent variable Y , i.e. the total number of transactions at lifetime point 12. In other words, for every user, you need to count how many transactions s/he had committed during her/his twelve months in Venmo.

We calculated dependent variable Y for each user. It is a constant throughout the users’ lifetime since it represents the total number of transactions during his/her twelve months in Venmo.



```
+----+-------------------+------------+---+-----------+
|user|           datetime|month_passed|  Y|transaction|
+----+-------------------+------------+---+-----------+
|   2|2012-11-22 22:03:42|           0|  1|          1|
|   3|2016-09-22 08:30:09|           0|  6|          1|
|   3|2016-10-08 18:56:24|           1|  6|          1|
|   3|2016-10-06 03:49:45|           1|  6|          1|
|   3|2016-10-07 16:37:56|           1|  6|          1|
|   3|2016-10-07 01:50:23|           1|  6|          1|
|   3|2016-10-08 20:36:13|           1|  6|          1|
|   4|2012-12-02 19:35:53|           0|  2|          1|
|   4|2012-12-14 21:51:12|           1|  2|          1|
|   8|2015-08-10 19:08:47|           0|  4|          1|
|   8|2016-03-01 23:35:34|           7|  4|          1|
|   8|2016-04-17 22:46:42|           9|  4|          1|
|   8|2016-05-16 18:18:05|          10|  4|          1|
|   9|2012-06-27 21:28:32|           0|  5|          1|
|   9|2012-08-12 23:00:53|           2|  5|          1|
|   9|2012-11-09 19:03:15|           5|  5|          1|
|   9|2012-12-28 00:52:01|           7|  5|          1|
|   9|2013-02-18 05:14:26|           8|  5|          1|
|  10|2012-11-25 01:20:39|           0|  8|          1|
|  10|2012-12-23 03:08:45|           1|  8|          1|
+----+-------------------+------------+---+-----------+
only showing top 20 rows
```

Q8: Create the recency and frequency variables. In CRM, this predictive framework is known as RFM. Here, you don’t have monetary amounts, so we will focus on just RF. Recency refers to the last time a user was active, and frequency is how often a user uses Venmo in a month. You need to compute these metrics across a user’s lifetime in Venmo (from 0 up to 12).

For example, if a user has used Venmo twice during her first month in Venmo with the second time being on day x, then her recency in month 1 is “30-x” and her frequency is 2/30=0.6667.

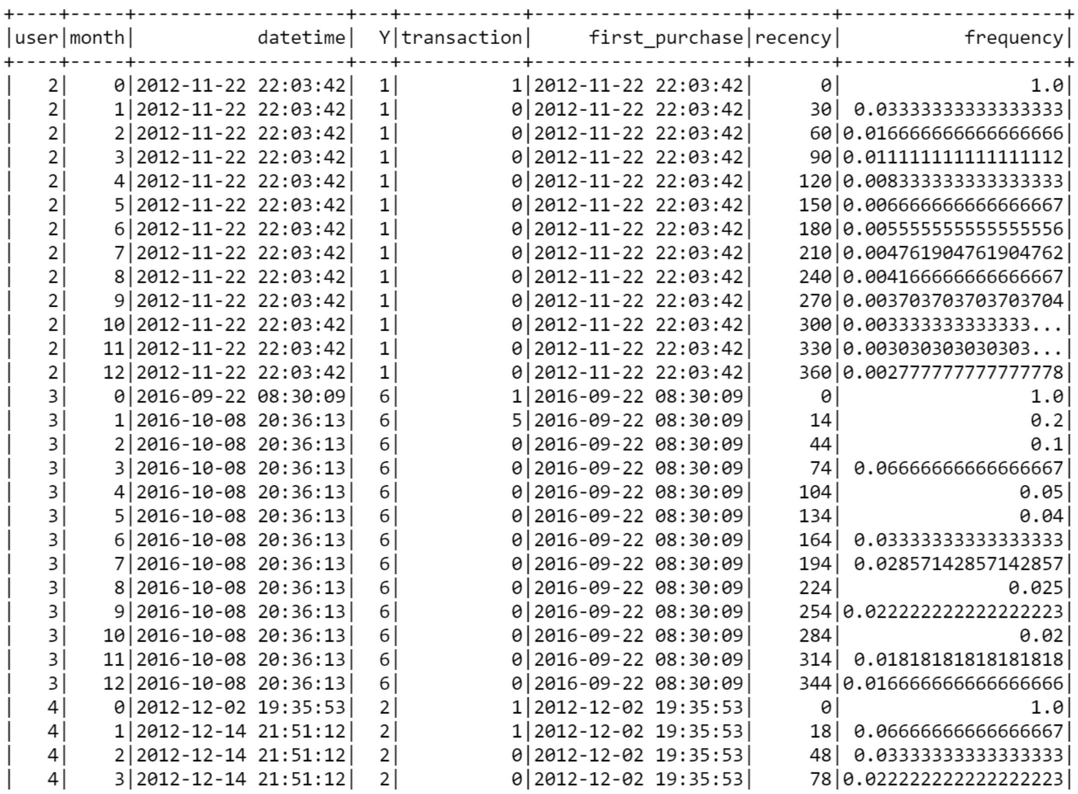
Here we calculate the recency and frequency metrics of each customer. It worth noticing that all time we use is relevant time. Just as mentioned previously in Question 4. (eg. Month 0 refers to the first transaction of the users, Month 1 refers to 30 days period after the users’ first transactions, etc.)

For Recency:

$$30 * month - (t_{Last\ time\ purchase} - t_{First\ time\ purchase})$$

For Frequency:

$$\frac{n_{transaction}}{30 * month}$$

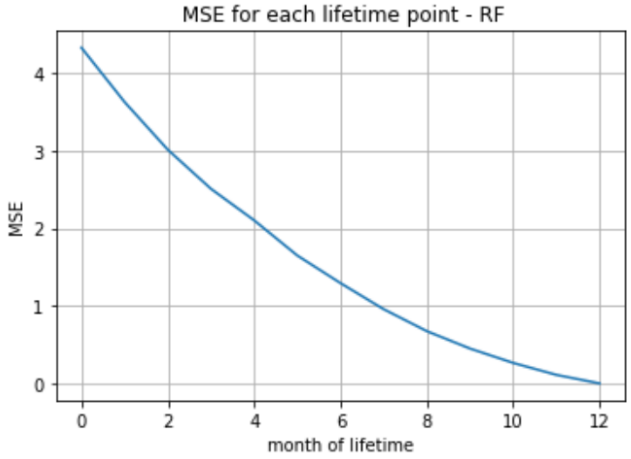


```
+----+-----+-------------------+---+-----------+-------------------+-------+--------------------+
|user|month|           datetime|  Y|transaction|     first_purchase|recency|           frequency|
+----+-----+-------------------+---+-----------+-------------------+-------+--------------------+
|   2|    0|2012-11-22 22:03:42|  1|          1|2012-11-22 22:03:42|      0|                 1.0|
|   2|    1|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|     30| 0.03333333333333333|
|   2|    2|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|     60|0.016666666666666666|
|   2|    3|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|     90|0.011111111111111112|
|   2|    4|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    120|0.008333333333333333|
|   2|    5|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    150|0.006666666666666667|
|   2|    6|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    180|0.005555555555555556|
|   2|    7|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    210|0.004761904761904762|
|   2|    8|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    240|0.004166666666666667|
|   2|    9|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    270|0.003703703703703704|
|   2|   10|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    300|0.003333333333333...|
|   2|   11|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    330|0.003030303030303...|
|   2|   12|2012-11-22 22:03:42|  1|          0|2012-11-22 22:03:42|    360|0.002777777777777778|
|   3|    0|2016-09-22 08:30:09|  6|          1|2016-09-22 08:30:09|      0|                 1.0|
|   3|    1|2016-10-08 20:36:13|  6|          5|2016-09-22 08:30:09|     14|                 0.2|
|   3|    2|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|     44|                 0.1|
|   3|    3|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|     74| 0.06666666666666667|
|   3|    4|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    104|                0.05|
|   3|    5|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    134|                0.04|
|   3|    6|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    164| 0.03333333333333333|
|   3|    7|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    194| 0.02857142857142857|
|   3|    8|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    224|               0.025|
|   3|    9|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    254|0.022222222222222223|
|   3|   10|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    284|                0.02|
|   3|   11|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    314| 0.01818181818181818|
|   3|   12|2016-10-08 20:36:13|  6|          0|2016-09-22 08:30:09|    344|0.016666666666666666|
|   4|    0|2012-12-02 19:35:53|  2|          1|2012-12-02 19:35:53|      0|                 1.0|
|   4|    1|2012-12-14 21:51:12|  2|          1|2012-12-02 19:35:53|     18| 0.06666666666666667|
|   4|    2|2012-12-14 21:51:12|  2|          0|2012-12-02 19:35:53|     48| 0.03333333333333333|
|   4|    3|2012-12-14 21:51:12|  2|          0|2012-12-02 19:35:53|     78|0.022222222222222223|
```

Q9: For each user's lifetime point, regress recency and frequency on Y. Plot the MSE for each lifetime point. In other words, your x-axis will be lifetime in months (0-12), and your y-axis will be the MSE. (Hint : Don't forget to split your data into train and test sets).

Here we split the datasets into train and test sets with ratio 0.7/0.3.

Then we split the datasets into month 0, month1, and so on. For each lifetime point, we regress recency and frequency on Y.

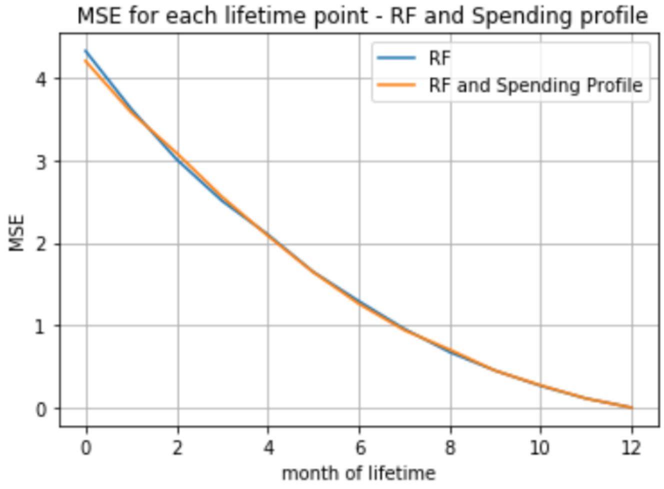


From the plot, we could see that MSE decreases as the lifetime goes on. It is reasonable because when we use the most recent data point, we could gain the information including all previous periods thus highly improving the prediction power.

Q10: For each user's lifetime point, regress recency, frequency AND her spending behavior profile on Y. Plot the MSE for each lifetime point like above. Did you get any improvement?

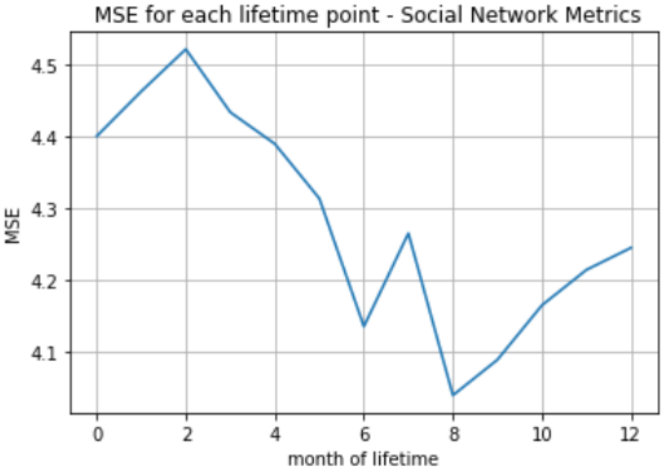
Yes, we get improvement.

The trend of the line is very similar. The MSE of the model adding spending behavior profile is smaller than the model with only RF. However, the difference here is very small and it looks like the two lines are overlapped.

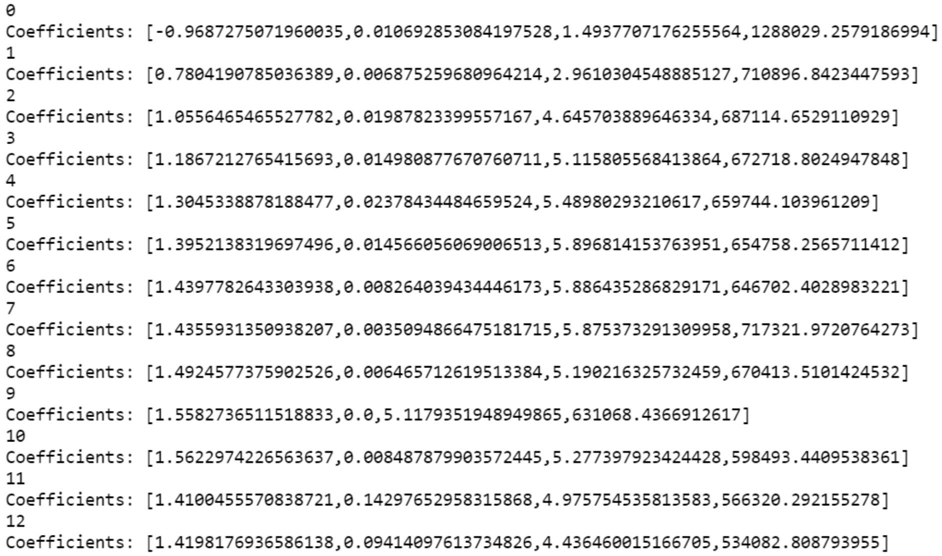


Q11: For each user's lifetime point, regress her social network metrics on Y. Plot the MSE for each lifetime point like above. What do you observe? How do social network metrics compare with the RF framework? What are the most informative predictors?

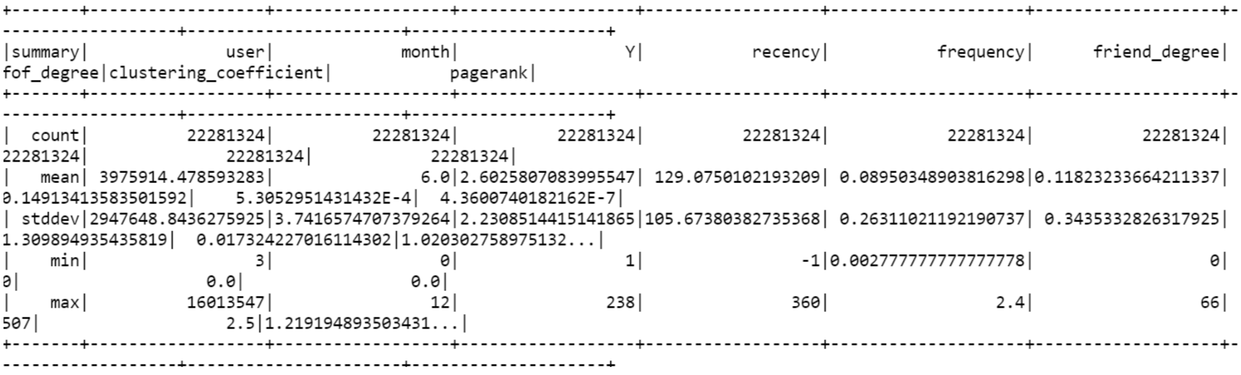
The trend is going down through the lifetime, but the line graph is a bit fluctuating here. As the MSEs for social network metrics are larger than that of the RF framework, RF framework outperforms the social network metrics here.



From the coefficients, we can see that the coefficient of the Pagerank has the biggest effect size, which indicates that Pagerank is the most informative predictors.



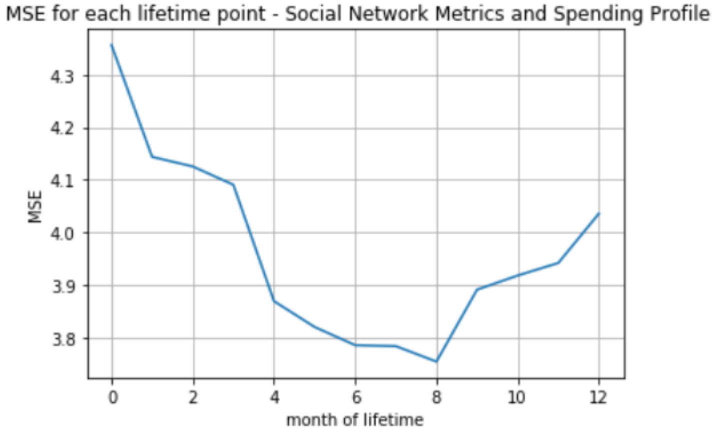
```
0
Coefficients: [-0.9687275071960035,0.010692853084197528,1.4937707176255564,1288029.2579186994]
1
Coefficients: [0.7804190785036389,0.006875259680964214,2.9610304548885127,710896.8423447593]
2
Coefficients: [1.0556465465527782,0.01987823399557167,4.645703889646334,687114.6529110929]
3
Coefficients: [1.1867212765415693,0.014980877670760711,5.115805568413864,672718.8024947848]
4
Coefficients: [1.3045338878188477,0.02378434484659524,5.48980293210617,659744.103961209]
5
Coefficients: [1.3952138319697496,0.014566056069006513,5.896814153763951,654758.2565711412]
6
Coefficients: [1.4397782643303938,0.008264039434446173,5.886435286829171,646702.4028983221]
7
Coefficients: [1.4355931350938207,0.0035094866475181715,5.875373291309958,717321.9720764273]
8
Coefficients: [1.4924577375902526,0.006465712619513384,5.190216325732459,670413.5101424532]
9
Coefficients: [1.5582736511518833,0.0,5.1179351948949865,631068.4366912617]
10
Coefficients: [1.5622974226563637,0.008487879903572445,5.277397923424428,598493.4409538361]
11
Coefficients: [1.4100455570838721,0.14297652958315868,4.975754535813583,566320.292155278]
12
Coefficients: [1.4198176936586138,0.09414097613734826,4.436460015166705,534082.808793955]
```



```
+-------+------------------+------------------+------------------+------------------+--------------------+-------------------+-
------------------+---------------------+--------------------+
|summary|              user|             month|                 Y|           recency|           frequency|      friend_degree|
fof_degree|clustering_coefficient|           pagerank|
+-------+------------------+------------------+------------------+------------------+--------------------+-------------------+-
------------------+---------------------+--------------------+
|  count|          22281324|          22281324|          22281324|          22281324|            22281324|           22281324|
22281324|              22281324|            22281324|
|   mean| 3975914.478593283|               6.0|2.6025807083995547| 129.0750102193209| 0.08950348903816298|0.11823233664211337|
0.14913413583501592|    5.3052951431432E-4|  4.3600740182162E-7|
| stddev|2947648.8436275925|3.7416574707379264|2.2308514415141865|105.67380382735368| 0.26311021192190737| 0.3435332826317925|
1.309894935435819|  0.017324227016114302|1.020302758975132...|
|    min|                 3|                 0|                 1|                -1|0.002777777777777778|                  0|
0|                   0.0|                 0.0|
|    max|          16013547|                12|               238|               360|                 2.4|                 66|
507|                   2.5|1.219194893503431...|
+-------+------------------+------------------+------------------+------------------+--------------------+-------------------+-
------------------+---------------------+--------------------+
```

Q12: For each user's lifetime point, regress her social network metrics and the spending behavior of her social network on Y. Plot the MSE for each lifetime point like above. Does the spending behavior of her social network add any predictive benefit compared to Q10?

After adding spending behavior, MSE plot is much smoother than before and MSEs decrease in each lifetime point, which indicates that spending behavior add some predictive benefit compared to previous question. However, the spike in the last month still exists, which still need to be explored in the future.


MSE for each lifetime point - Social Network Metrics and Spending Profile
MSE
month of lifetime