Showwcase Data Analyst Intern: Data Challenge

Name: Akshay Kajale
Computer Science Graduate Student
Sab Jose State University
Email: akshay.kajale@sjsu.edu
Mob No: (669)-278-8855

Challenge

The product team at Showwcase wants to better understand "**user engagement**" on the platform in the last month (October 2019 in this exercise). You are free to define engagement, find insights, and question any other metrics we can use to understand our approach better. Your role is to present your findings based on the data.

Here are some of my findings:

Preprocessing:

1. To study about the individual user engagement, I aggregated the data (group by) by customer_id and removed all the Boolean variables.
2. The aim behind this aggregation was to get total projects, likes, comments, session duration, bugs faced etc by individual user.

```
In [10]: runfile('/Users/akshay9/E_D_A.py', wdir='/Users/akshay9')
    customer_id  Total_Projects_Added  Total_bugs_faced_by_User  Total_Duration
0         98346                     0                         5             103
1         92736                     3                         4             127
2         23457                     2                         0             207
3         20548                     3                         1             258
4         43658                     3                         0             522
5         23596                     2                         0             640
6         68754                     3                         0             822
7         93486                     2                         0             872
8         33496                     2                         0            1154
9         43096                     1                         0            1172
10        29345                     0                         0            1291
11        46329                     3                         0            1299
12        23548                     0                         0            1327
13        19348                     1                         0            1565
14        19374                     1                         0            1633
15        82963                     3                         0            1869
16        24520                     3                         5            1957
17        98476                     2                         0            2051
18        76530                     0                         2            2091
19        10246                     8                         9            2863
20        39475                     9                        14            4276
21        93464                     7                         6            5348
22        73524                    23                         9            5448
23        80746                     8                         8            5671
24        23579                     8                         5            5718
25        98653                    12                         5            5747
26        43086                     7                         5            6221
27        12496                     9                         4            6279
28        43549                    10                         7            6338
29        12940                     2                         3            6421
30        29356                     7                        11            6423
31        32047                     3                         0            7287
32        43057                     9                         8            7984
33        96254                    10                        10            8342
34        34574                    19                        11            8544
35        23083                    12                        16           12135
36        12407                    17                         6           12266
37        51243                    24                        14           12556
38        23985                    11                         8           13014
39        87265                    24                        12           16487
40        73245                    24                        24           17375
41        14354                    31                        15           17437
42        40572                    27                        25           17773
43        87323                    38                        25           17954
44        40235                    23                        19           21117
45        38459                    25                        15           22880
46        23404                    25                        33           24809
47        29375                    20                        26           30356

IPython console  History
```

Fig 1: Aggregated Data on Customer ID Sorted on Session Duration

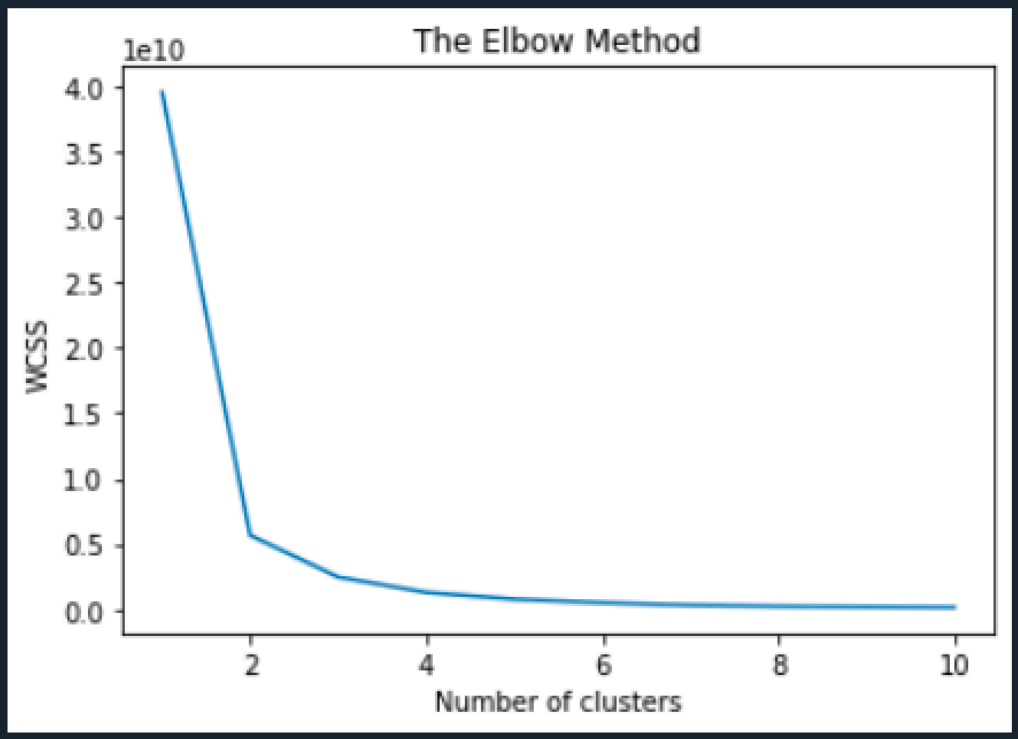
Here you can observe in the table that projects added, bugs and session have an increasing trend and are correlated with each other in a positive way. We can get an insight that the more the user is uploading the project, more it is facing the bugs are more is the session time. It can be one of the possibilities that user might move

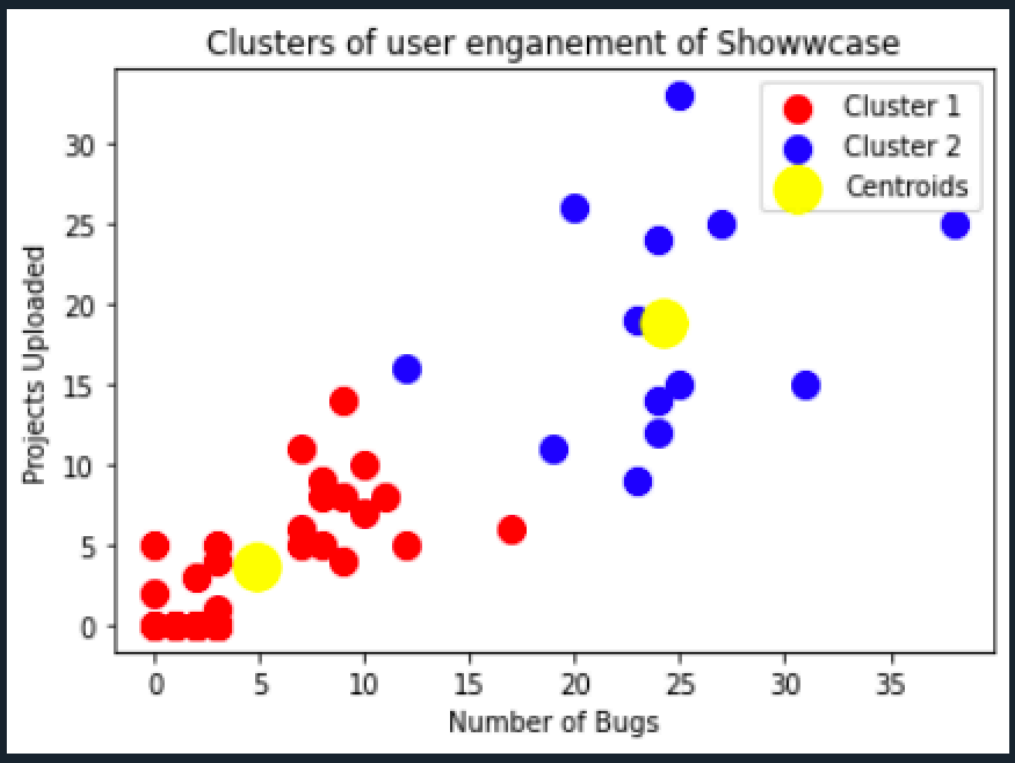
away from the platform. Specifically, if you see user 98346, 92736 and 39475, they have encountered comparatively a greater number of bugs as compared to total duration. That can be one of the reasons that people might move away from the Showwcase. We have to make sure to take care of these bugs to improve user experience which would ultimately attract more users towards platform.

Attribute selection output

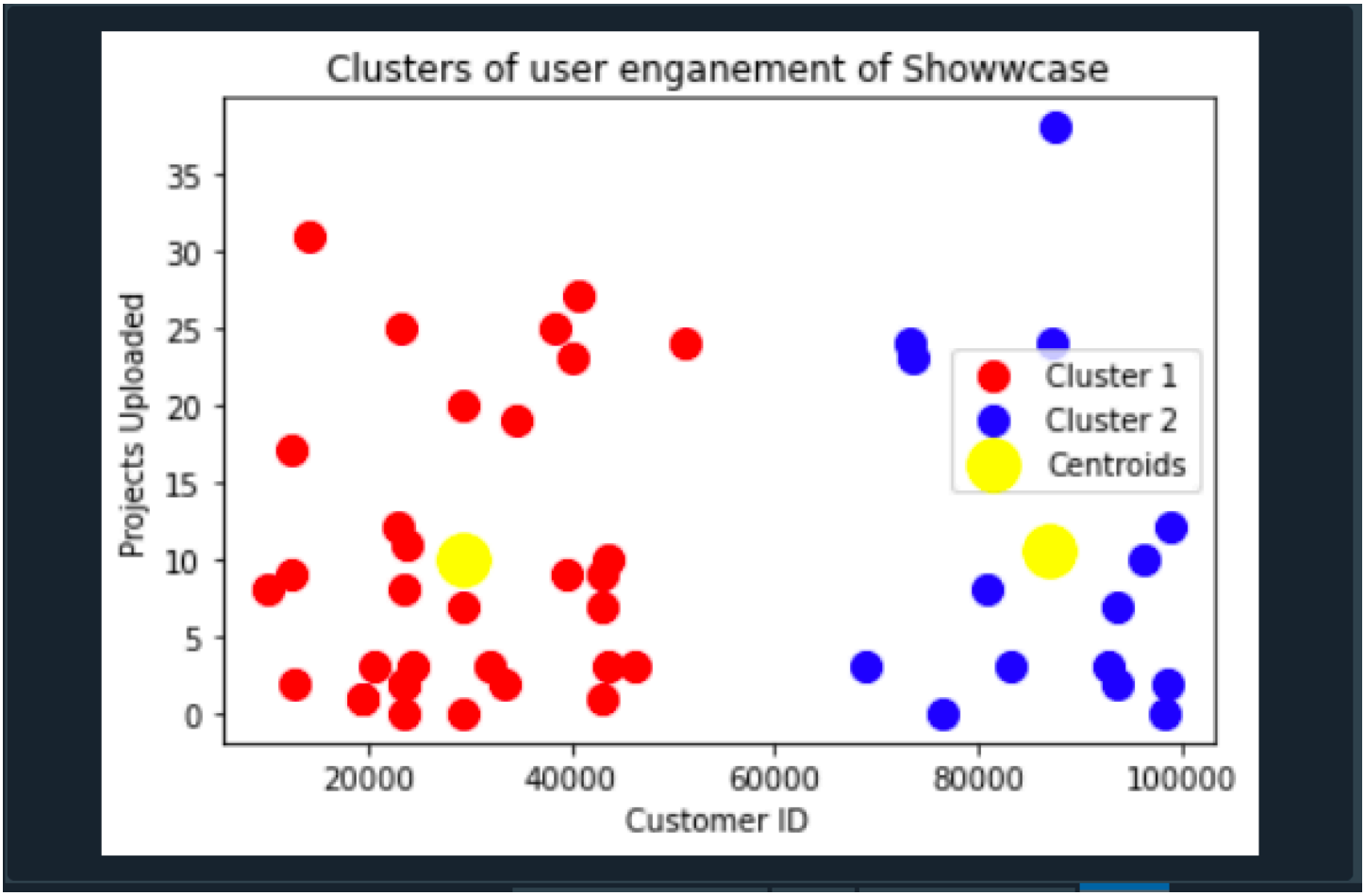
```
=== Run information ===

Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     showwcase_data
Instances:    300
Attributes:   14
               session_id
              customer_id
              login_date
              projects_added
              likes_given
              comment_given
              inactive_status
              bug_occured
              session_projects_added
              session_likes_given
              session_comments_given
              inactive_duration
              bugs_in_session
              session_duration
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 14 session_duration):
        Correlation Ranking Filter
Ranked attributes:
 0.0708  12 inactive_duration
 0.0679   7 inactive_status
 0.0404   3 login_date
 0.0391   4 projects_added
 0.0265   5 likes_given
 0.0254   8 bug_occured
 0.0181   6 comment_given
 0.0168   9 session_projects_added
 0.0115  11 session_comments_given
-0.0134   1  session_id
-0.0291  13 bugs_in_session
-0.0523   2 customer_id
-0.0568  10 session_likes_given

Selected attributes: 12,7,3,4,5,8,6,9,11,1,13,2,10 : 13
```

Here I performed some exploratory data analysis on attributes. The dependent attribute was session duration and you can see that inactive_duration is positively correlated with session duration. It means that as the session duration is increasing, the inactive duration is also increasing which is not a good sign. There might be several possibilities why inactive duration is increasing. Maybe user is facing more bugs that's why inactive duration is high. You can also see that bug_occured is also positively correlated with session_duration which means it is also correlated with inactive_duration.

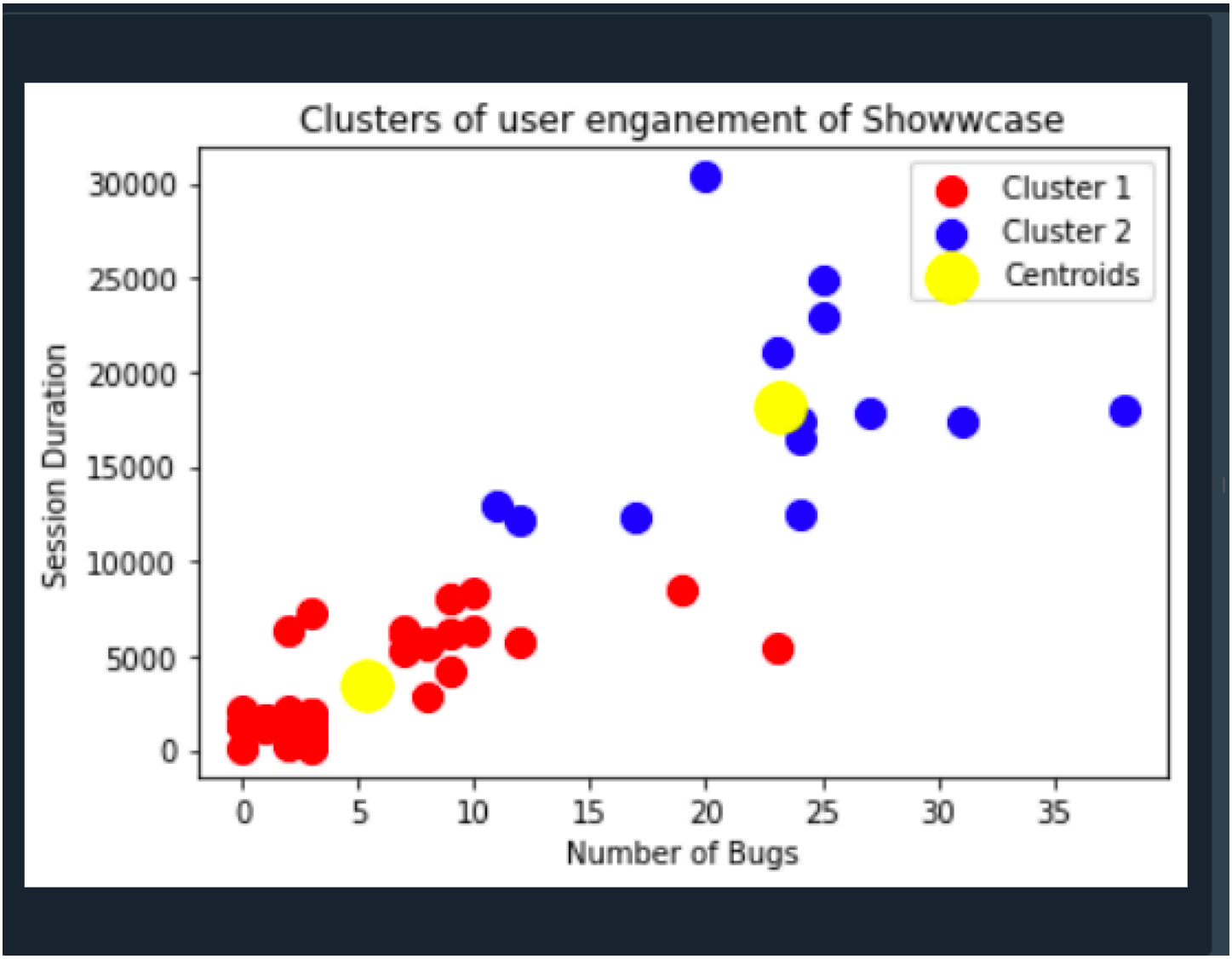




Here in above image, I performed K-means clustering to classify Projects uploaded vs Number of Bugs in two clusters. Here you can see the elbow curve to decide the optimal number of clusters (which are 2). Here you can observe the trend that people who are encounter more bugs are uploading more projects. But at the same time, according to first table, it is taking a lot of time to upload these projects.



Here is just a clustering representation of projects uploaded vs Customer ID. Here we can observe that ID's in the range of 0-40000 are more active as they have as they have a greater number of people who have uploaded greater than 10 projects but cluster 2 (70000-100000) have uploaded a smaller number of projects. I this part, I think we have to concentrate on the people who are uploading a smaller number of projects (below 10 or 5) and try to find if they are facing any issues. Because number of projects uploaded is directly proportional to total duration they spend on the platform.



This is a cluster representation of Total session duration vs Number of bugs. Here if you observe as the session duration increases, number of bugs also increases. This should not happen as it would not be a good User Experience. This could tend away the users.

Here is my analysis of the data you provided. If given an opportunity, I would really love to discuss and explain the findings in person. Thank you for the opportunity.