

# Analysis on Showcase data

Tony Nguyen

09/17/2020

# Table of content

<b>1</b>	<b>Abstract</b>	<b>3</b>
<b>2</b>	<b>Generating a dataframe with the csv using python</b>	<b>4</b>
2.1	Building a duration column and tools . . . . .	4
2.2	Calculating the duration time . . . . .	5
<b>3</b>	<b>Dicussing the features</b>	<b>6</b>
3.1	Discussing the given bar graphs . . . . .	7
<b>4</b>	<b>Looking at bar graphs with certain conditions</b>	<b>10</b>
4.1	Bar graphs with 0 bugs . . . . .	13
4.2	Bar graphs with 0 likes . . . . .	16
<b>5</b>	<b>Future works</b>	<b>19</b>
5.1	Conclusion and other considerations . . . . .	19

# 1 Abstract

The goal of this project is : given a csv file that contains about 300 data session id, customer id, etc find out what features show user engagement. To start, we first import the csv into jupyter notebook to generate a dataframe(using pandas) that can be easier to read. Next we need to define user engagement, user engagement is the concept where users are constantly interacting with some system(in this case showcase website). From this definition we could argue that user engagement can be either working with bugs, liking comments, or adding more projects, however, factors such as liking may not keep a person on the website for a long time. To find out what is the duration of a user in a session, this can be done by subtracting the duration time to idle time to show us how much time a person has actually spent online. After generating a duration column, we then generate bar graphs to find out the days users spent on the website performing tasks such as liking a project, adding a project, encountering bugs. With generating graphs, the features that aren't important would be the the numbers of likes given on a certain post and the number of bugs encounter in a session. The reason likes don't show much user engagement is because liking posts are short interactions, they don't keep the person on long enough, bugs are user engagement but they aren't the reason to keep someone on.

## 2 Generating a dataframe with the csv using python

### 2.1 Building a duration column and tools

To build the dataframe, we first use jupyter notebook. Next we need to import two things pandas and csv. We can also import matplotlib early so that we can use it later to build our graphs. To build the dataframe we need to then make the dataframe read the csv. In this case the pathing to the csv is on desktop to make the pathing easier. Figure 2 displays the output of the first 10 dataset due to the df.head.

```
import pandas as pd
import csv
import matplotlib.pyplot as plt

df=pd.read_csv("\\Users\\Administrator\\Desktop\\showcase_sessions.csv")
df.head(10)
```

Figure 1: The code for importing the csv file

	session_id	customer_id	login_date	projects_added	likes_given	comment_given	inactive_status	bug_occured	session_projects_added	session_likes_given
0	624205.0	80746.0	10/30/19	False	True	True	True	False	0.0	24.0
1	624241.0	24520.0	10/30/19	True	True	True	True	False	2.0	3.0
2	111002.0	32047.0	10/30/19	True	True	True	True	False	1.0	5.0
3	545113.0	23404.0	10/30/19	True	True	True	False	False	1.0	10.0
4	750269.0	40235.0	10/30/19	True	True	False	True	False	3.0	16.0
5	744943.0	73245.0	10/30/19	True	True	True	True	True	3.0	27.0
6	922001.0	12407.0	10/30/19	True	False	True	False	False	5.0	0.0
7	823895.0	29375.0	10/30/19	False	False	True	True	False	0.0	0.0
8	490096.0	40572.0	10/30/19	True	True	False	False	False	1.0	25.0
9	919319.0	23404.0	10/29/19	True	True	False	True	False	2.0	14.0

Figure 2: output of the first 10 data

## 2.2 Calculating the duration time

Next we need to build a new column that shows duration time. The reason for this is because we are given two types of duration: inactive and session. Inactive is the time you are idle while being on the website compare to session which is the total time being on the website. These two do not tell us how long the person actually been on performing a task, they only tell us either how long the person is away or the total time the person was using the website. To make this more useful, we create a column call duration time that calculate the difference of inactive and session. This is shown in figure 3 and figure 4 shows the output of the new column with the duration time.

```
df['duration_time'] = df['session_duration'] - df['inactive_duration']
neg_duration = df[df['duration_time'] < 0].index
df.drop(neg_duration, inplace=True)
df.head(10)
```

Figure 3: calculation of duration

status	bug_occured	session_projects_added	session_likes_given	session_comments_given	inactive_duration	bugs_in_session	session_duration	duration_time
True	False	0.0	24.0	3.0	1146.0	0.0	1564.0	418.0
True	False	2.0	3.0	5.0	133.0	0.0	1766.0	1633.0
True	False	1.0	5.0	5.0	1571.0	0.0	2230.0	659.0
False	False	1.0	10.0	21.0	0.0	0.0	633.0	633.0
True	False	3.0	16.0	0.0	1405.0	0.0	1679.0	274.0
False	False	5.0	0.0	5.0	0.0	0.0	1329.0	1329.0
False	False	1.0	25.0	0.0	0.0	0.0	290.0	290.0
True	False	0.0	0.0	1.0	930.0	0.0	1374.0	444.0
True	False	2.0	4.0	1.0	672.0	0.0	1138.0	466.0
False	False	3.0	24.0	0.0	0.0	0.0	39.0	39.0

Figure 4: output of the column

### 3 Dicussing the features

With this we can now look into what features would be useful or impactful towards user engagement. To reiterate, user engagement is where users are constantly interacting with some system. The features that are in this dataframe are: bugs in session, likes for projects, number of projects added, and number of comments given. We will first look at the duration time for each of these and see what trends we can learn from this. Figure 5 shows the code to building the bar graph with labels

```
plt.bar(df['session_projects_added'],df['duration_time'])
plt.xlabel('session_projects_added')
plt.ylabel('duration')
plt.show()
plt.bar(df['login_date'],df['duration_time'])
plt.xticks(rotation=60)
plt.xlabel('login_date')
plt.ylabel('duration')
plt.show()
plt.bar(df['session_likes_given'],df['duration_time'])
plt.xlabel('session_likes_given')
plt.ylabel('duration')
plt.show()
```

Figure 5: code for generating the bar graph

### 3.1 Discussing the given bar graphs

To start, figures 6 to 8 are bar plots of projects added, dates, and likes over certain durations. From figure 7, there could be certain times where people are on more often and on other days less. What we can look further for login date is what people are doing within those times or what would be consider popular to do. Figure 6 shows us the number of projects added, from this figure the common amount of projects added are around 0 to 3 anything more would mean less time involved. For figure 8 it shows number of likes given over certain duration. It doesn't really tell us what type of topics are being liked or if they are a good indicator of user engagement.

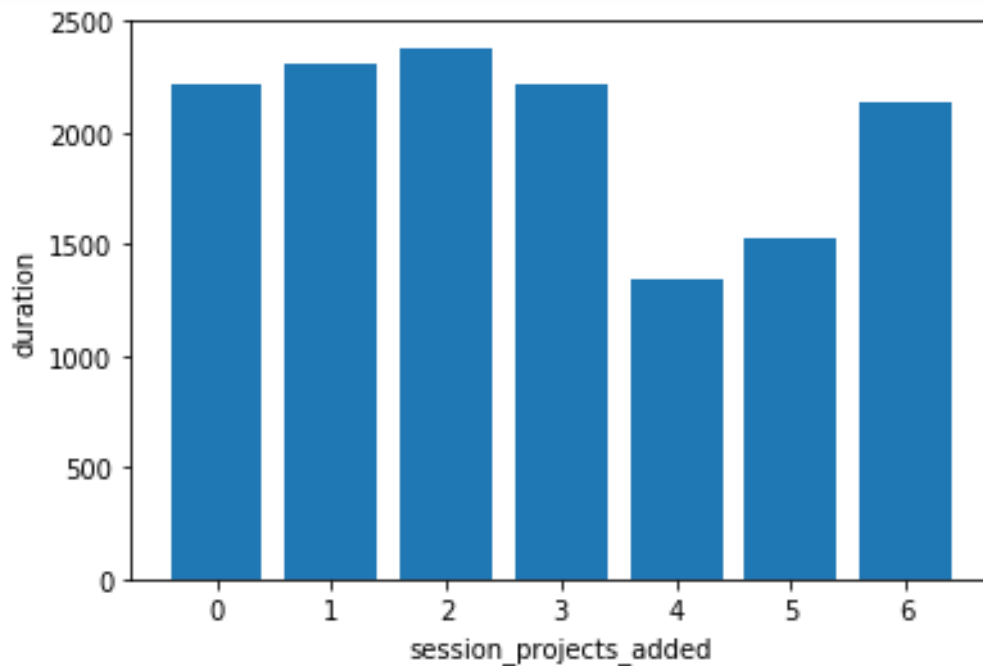


Figure 6: output of projects added over certain duration

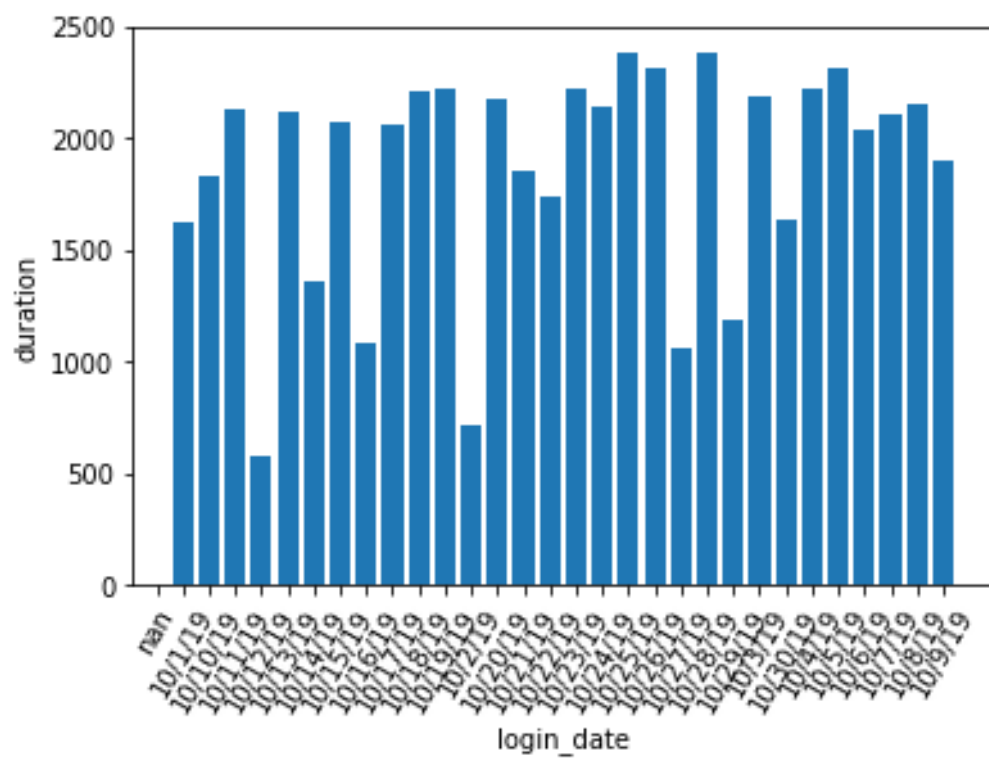


Figure 7: number of people logging in over certain duration



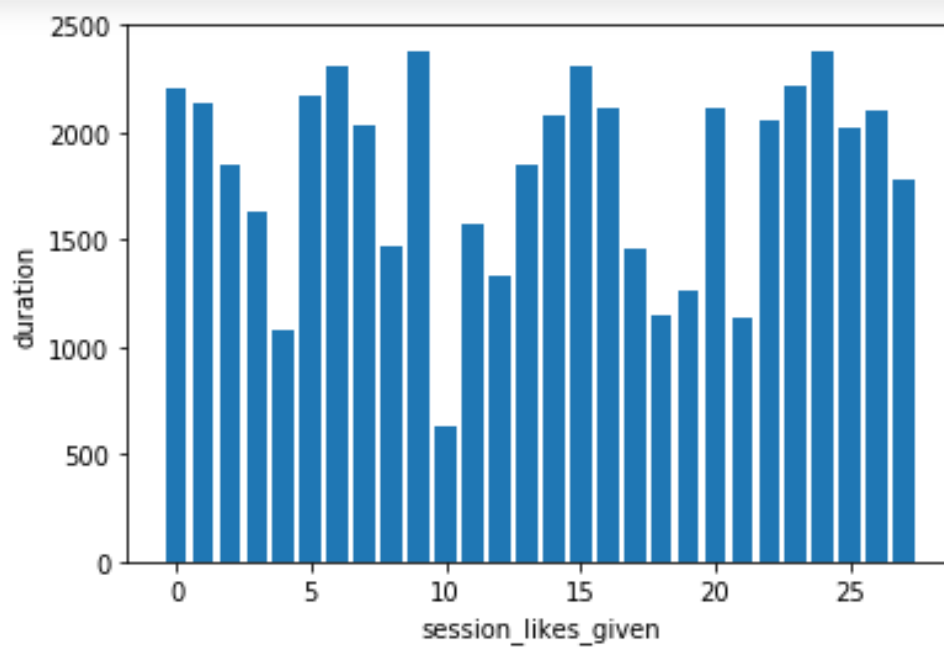


Figure 8: number of likes over certain duration

## 4 Looking at bar graphs with certain conditions

To find out which features are important we can look at how users work if with certain features removed. For the first case we will look at features with 0 projects added and see if these features do anything to the duration. For figure 9 the number of likes given during a session is still high despite 0 projects. That means that users who have added no projects are still providing likes toward others. For figure 10 comments are still being made by those who have 0 projects and the most people tend to make are 1 or 5 comments. For figure 11 most people with 0 projects have either no bugs or 2. This shows that the bugs that they are experiencing are possibly from visiting other parts of the site as they haven't added any projects.

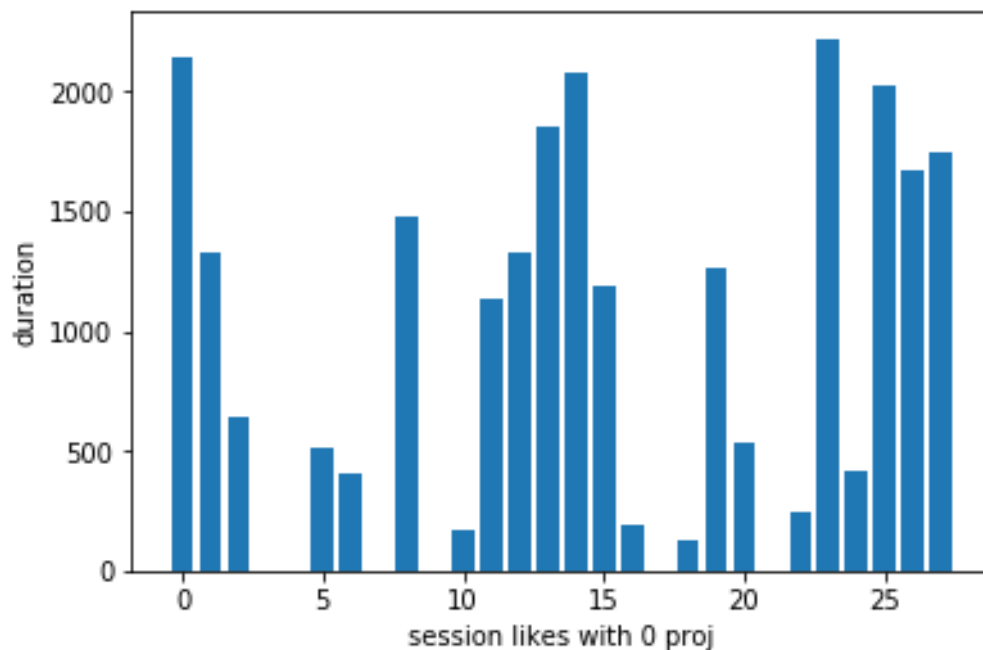


Figure 9: number of likes for those with 0 project

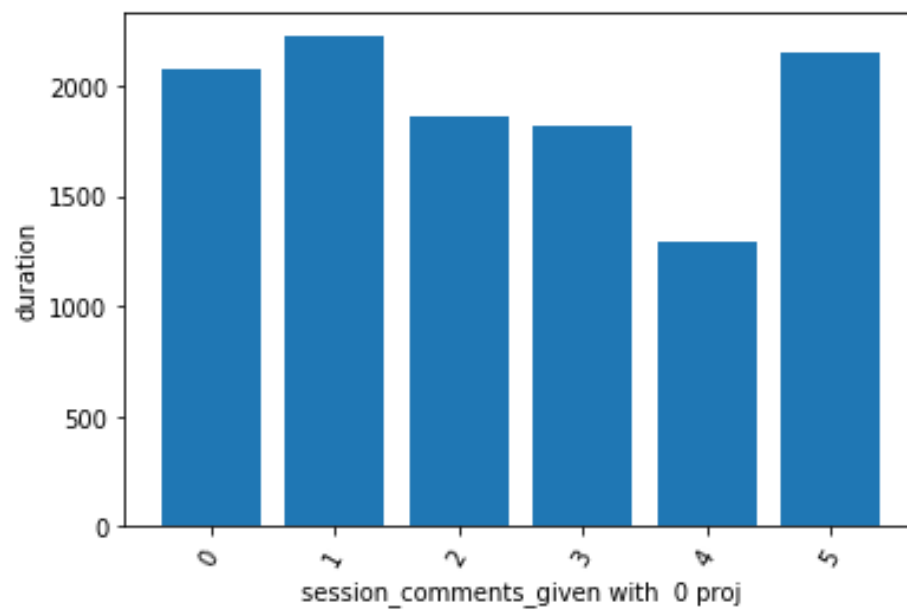


Figure 10: comments with 0 projects

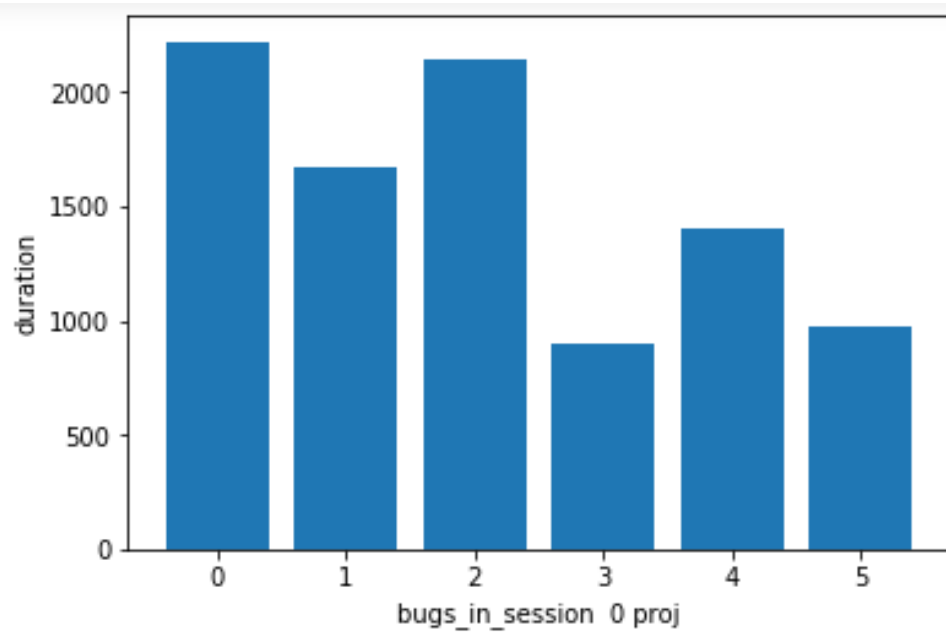


Figure 11: bugs in session 0 project

## 4.1 Bar graphs with 0 bugs

Looking at figure 12 the number of comments made by people who experience no bugs. People with no bugs seem to be providing 0 to 5 or 10 comments during a session. What this shows is people seem to be providing a consistent number of comments 0 to 5 during a session. Figure 13 shows the number of likes given by people who experience no bugs. What is interesting is people with 0 bugs will provide likes this doesn't really show us much as people can be focusing on other things while liking. Figure 14 shows us that people will add projects more when they have no bug experience and would be around longer. What can be understood from the case with 0 bugs is that people will tend to add more projects and liking during a session does not seem to change.

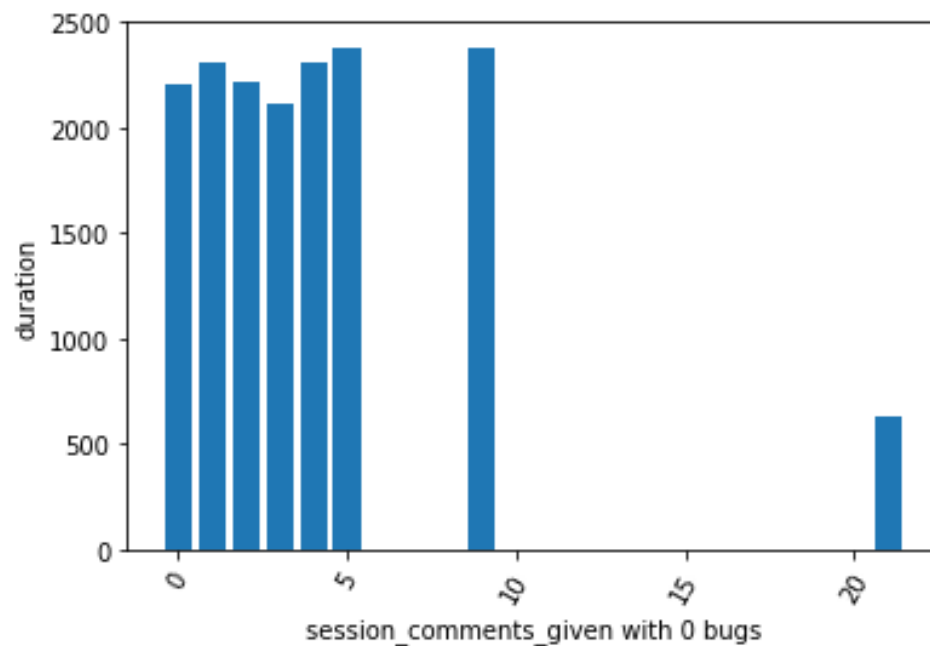


Figure 12: comments with 0 bugs

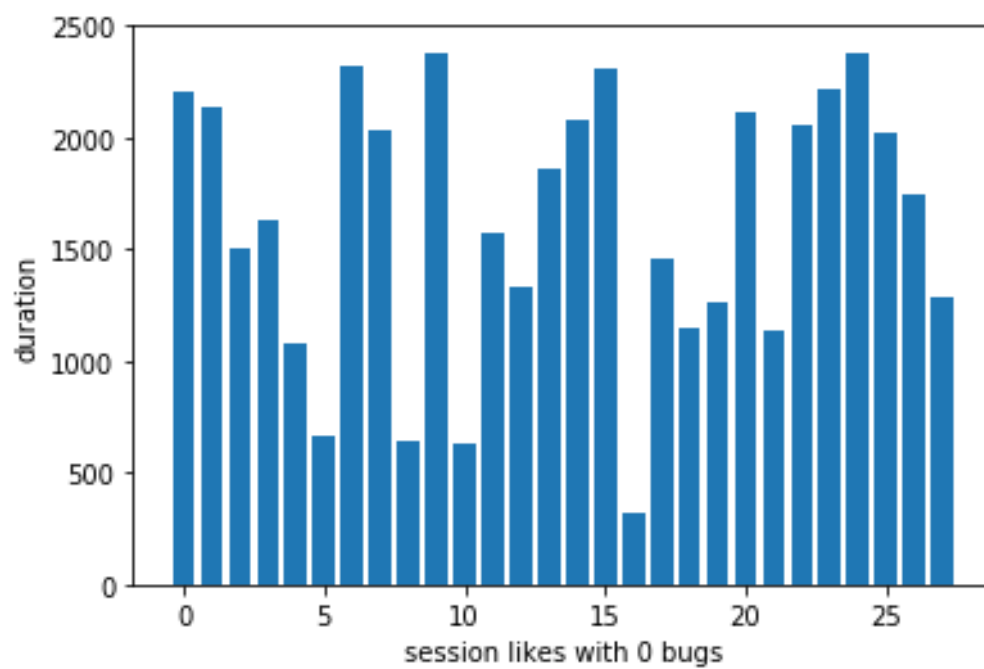


Figure 13: likes with 0 bugs

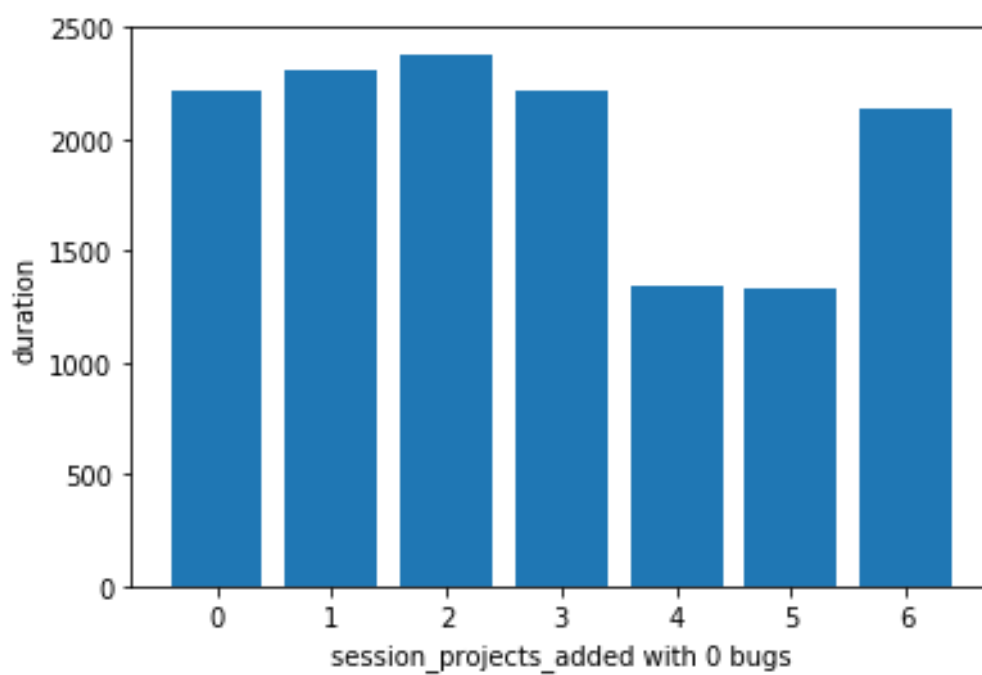


Figure 14: projects added with 0 bugs

## 4.2 Bar graphs with 0 likes

If we look at people who provides no likes they tend to either give no comments or 1 or 5 comments shown in figure 15. What this shows is the interaction is not dependent on the number of likes in a session. In figure 16, the number of bugs encounter by people who gives no likes are lowered. These users also have a tendency to add projects from 0 to 3.

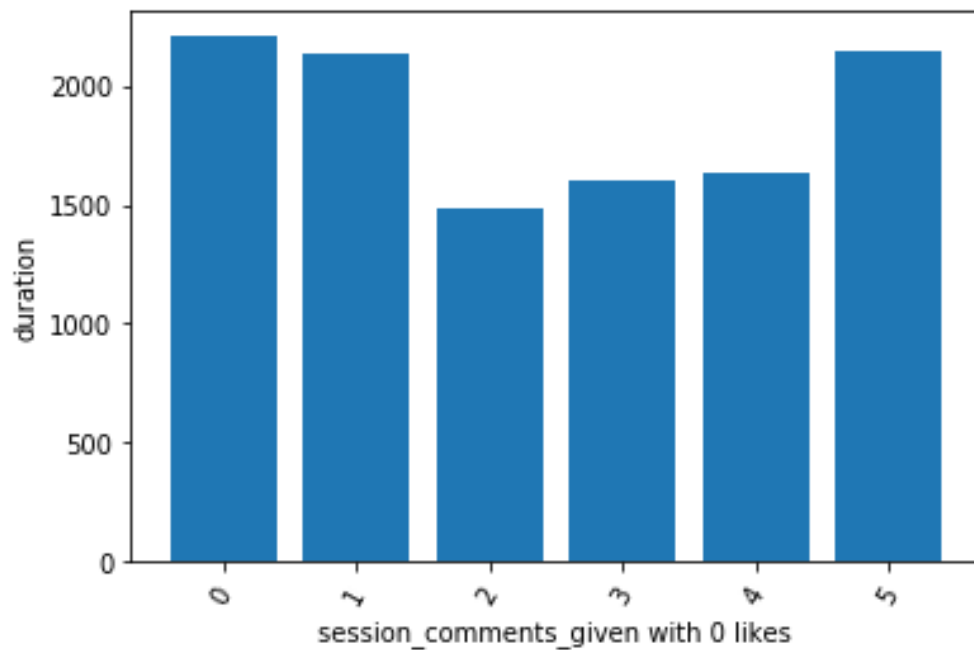


Figure 15: comments with 0 likes



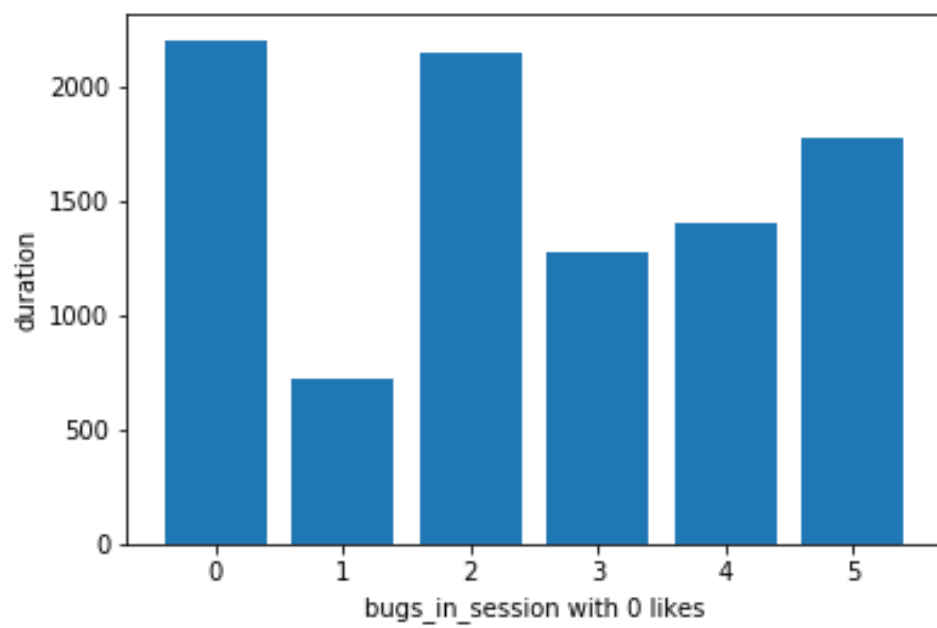


Figure 16: bugs in session with 0 likes

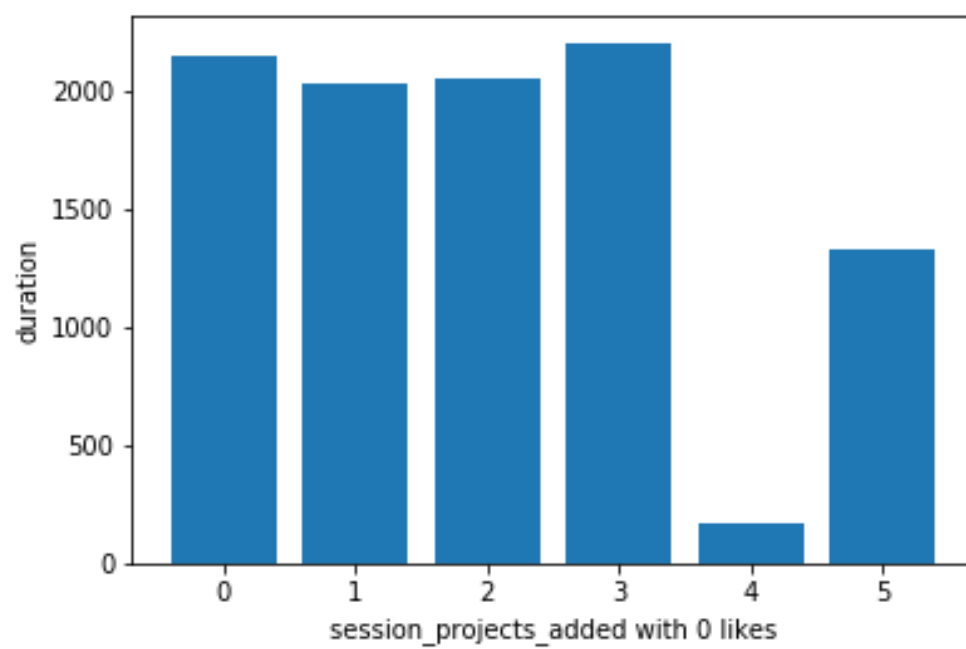


Figure 17: projects added 0 likes

## 5 Future works

### 5.1 Conclusion and other considerations

What could be found from the data is that features we can use as user engagements are the number of comments people make and number of projects added. Based off the graphs the most people are on are mostly from adding new projects or discussing on other people's work. What we can also look more into is finding out what types of comments are being made and what type of projects are being put on the site. Knowing this, we can find a trend on what types of projects are mostly keeping users interacting and finding out what is the common size of the comment. Future work can be we can look more into user retention and find out what is keeping a user coming back and interacting on the website longer and see if there are other possible features that can be discovered. Another interesting concept to work with is applying machine learning to see what feature does the computer find important for us and what makes a user stay on longer.