# Modeling User Engagement

Steven Nguyen

September 17, 2020

# Table of content

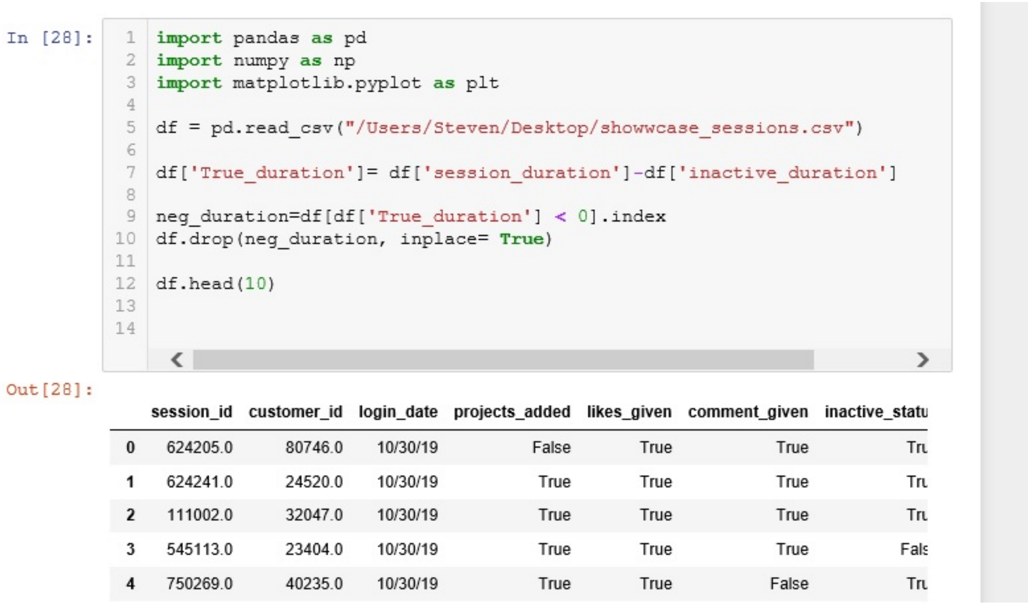
# 1 Abstract

We are given 300 CSV files from ShowCase with information( such as how long a session is, customer id, number of bugs encountered, projects added, likes given, comments given), and the goal is to use this information to find trends that can help us define user engagement. To accomplish this work, I will be using Python and its packages such as pandas, numpy, and matplotlib to graph the data given and find trends that will help us better define user engagement. Based off the information provided, my hypothesis is user engagement can be determined by two variables: number of projects added, and the number of comments given. We can use this justification to define user engagement as the interaction between Showcase's customer and its product. While other categorizes such as number of bugs encountered and likes given may be prevalanet in all forms of data, I believe they do not defined user engagement. Number of bugs and number of likes given can however show us other useful information such as which category tends to have more bugs or users using the like feature to help friends projects gain relevancy.

# 2 Required Packages and Code Discussion

## 2.1 Packages and Code Discussion

The coding language I will be using is Python and the packages required will be Pandas, Numpy, and Matplotlib. I begin by uploading the CSV file and naming this CSV file as my own dataframe. From there I subtracted the session's duration from the idle duration to get the true duration a user is online. I have done this because I have seen some of the files containing idle duration to be significantly higher than the session's duration( which implies a user is idle longer than the total time he spent on the session). I use the Panda's drop to drop all negative true duration and focus mainly on positive values.



In [28]:

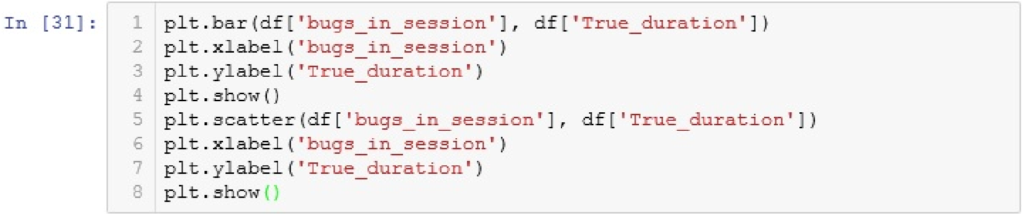
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("/Users/Steven/Desktop/showcase_sessions.csv")

df['True_duration']= df['session_duration']-df['inactive_duration']

neg_duration=df[df['True_duration'] < 0].index
df.drop(neg_duration, inplace= True)

df.head(10)
```

Out[28]:

```
   session_id  customer_id  login_date  projects_added  likes_given  comment_given  inactive_statu
0    624205.0      80746.0    10/30/19           False         True           True             Tru
1    624241.0      24520.0    10/30/19            True         True           True             Tru
2    111002.0      32047.0    10/30/19            True         True           True             Tru
3    545113.0      23404.0    10/30/19            True         True           True             Fals
4    750269.0      40235.0    10/30/19            True         True          False             Tru
```

Figure 1: Pandas drop feature to eliminate negative True Duration

From here, I attempt to find trends using the features: number of bugs, projects added per session, likes given per session, and comments given per session and plot them against the True duration. I plot these by using Matplotlib and its "bar" graph feature. I will also be using scatter plots in this entire experiement because I believe while scatter plots will become bar plots over a large number of datasets, I would like to see if I can extrapolate any other additional information. Below is the code and the figures of each variable.



In [31]:
```
plt.bar(df['bugs_in_session'], df['True_duration'])
plt.xlabel('bugs_in_session')
plt.ylabel('True_duration')
plt.show()
plt.scatter(df['bugs_in_session'], df['True_duration'])
plt.xlabel('bugs_in_session')
plt.ylabel('True_duration')
plt.show()
```

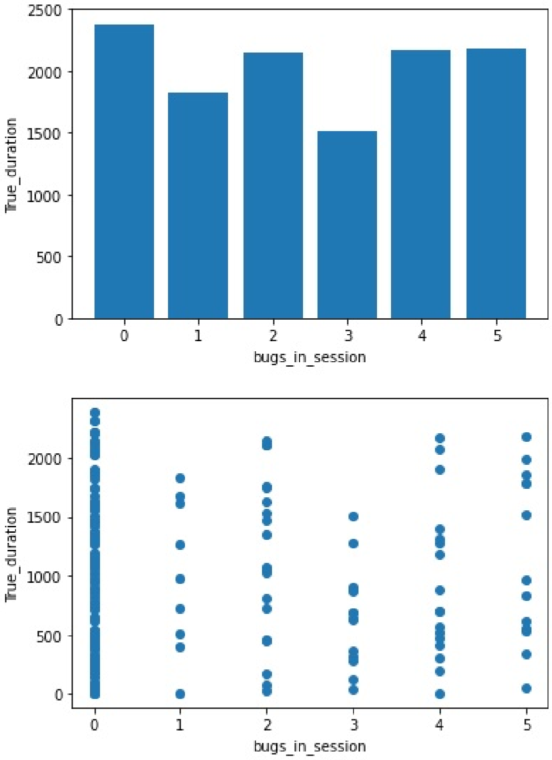
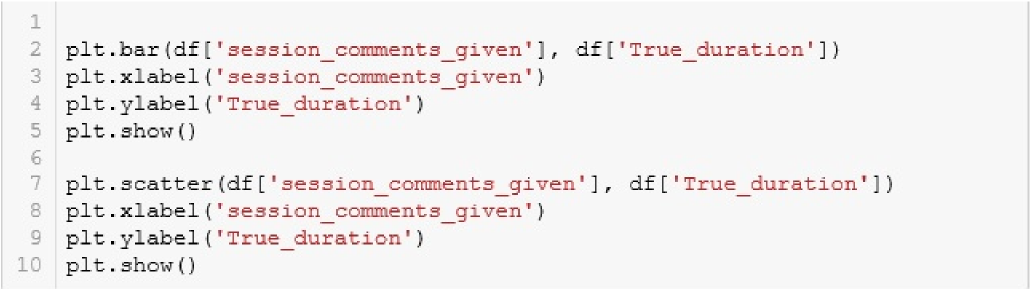


Figure 2: Figure of both Scatter and Bar graph for bugs in session vs True duration

This figure (the number of bugs) informs us that generally users will experience no bugs, based off the scatter plot with the 0 bar. However those that do experience bugs will experience them at varying intensity and duration. This figure still may not be enough information to prove if it is a variable for user engagement.



```
plt.bar(df['session_comments_given'], df['True_duration'])
plt.xlabel('session_comments_given')
plt.ylabel('True_duration')
plt.show()

plt.scatter(df['session_comments_given'], df['True_duration'])
plt.xlabel('session_comments_given')
plt.ylabel('True_duration')
plt.show()
```

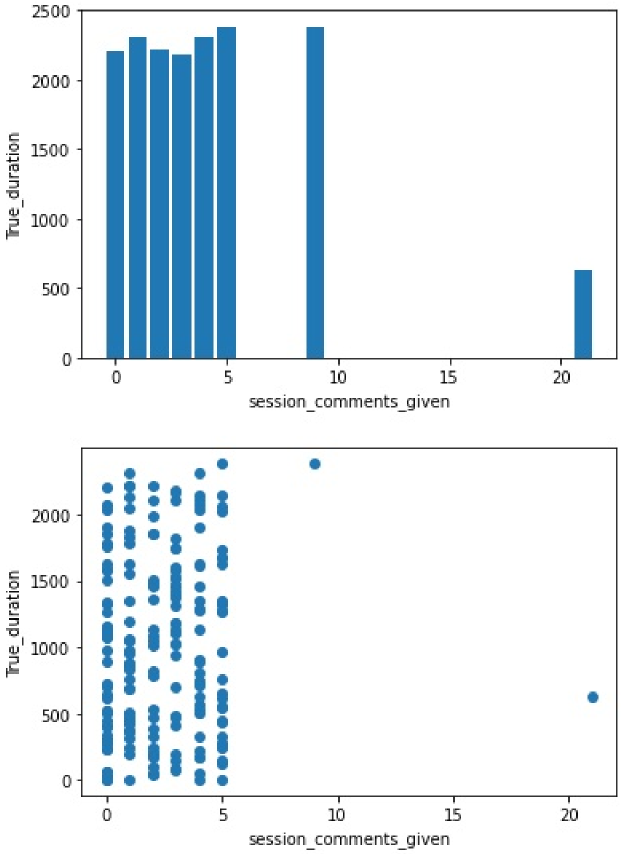
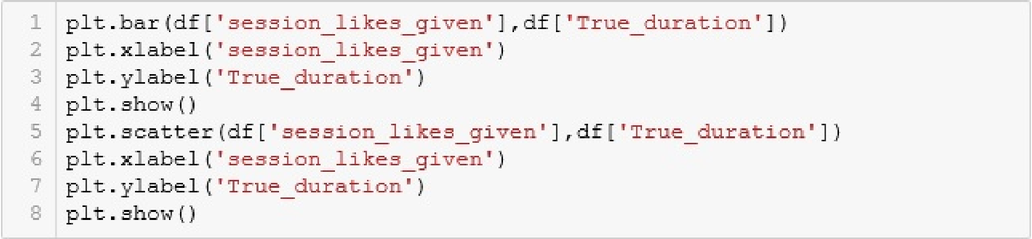


Figure 3: Figure of Scatter and Bar graph of Session Comments vs True duration

This figure tells us the behavoir of the customers using Showcase. Generally, customers prefer to post up to the maximum of 5 comments in a session (with a couple of outliers outside of that range). Like the bugs figure, this may not be sufficient information to prove user engagement. This figure of



```python
plt.bar(df['session_likes_given'],df['True_duration'])
plt.xlabel('session_likes_given')
plt.ylabel('True_duration')
plt.show()
plt.scatter(df['session_likes_given'],df['True_duration'])
plt.xlabel('session_likes_given')
plt.ylabel('True_duration')
plt.show()
```

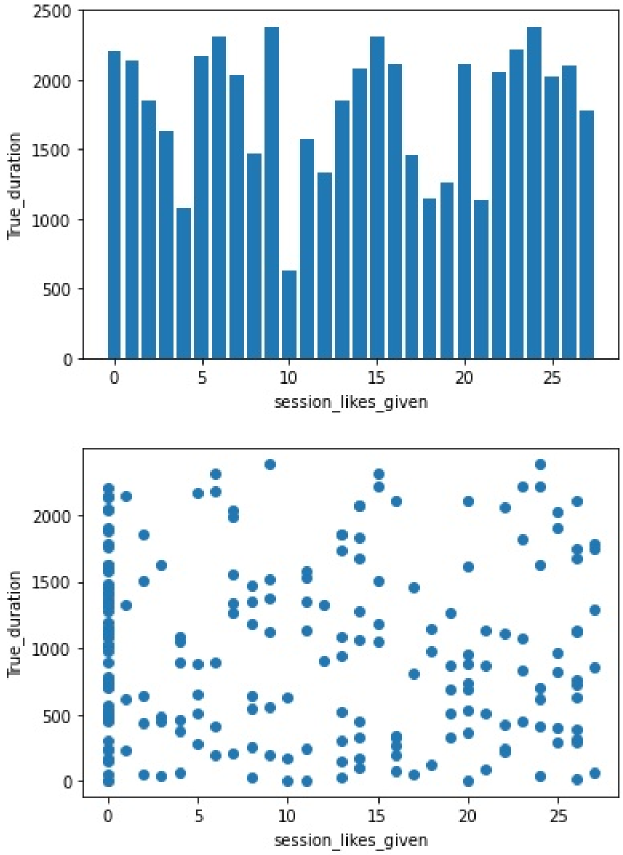
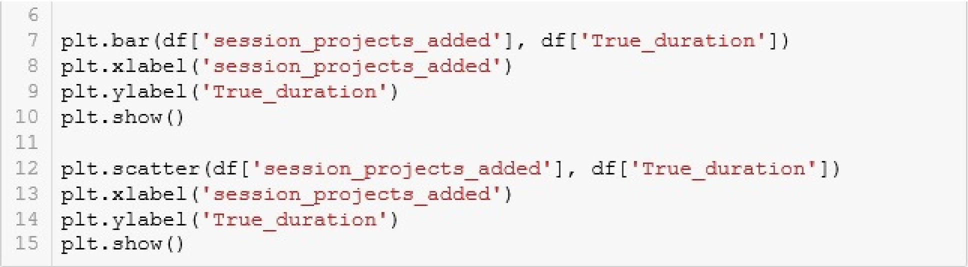


Figure 4: Figure of Session Likes vs True duration in both Scatter and Bar graphs

session likes shows us customers typically give likes on different timeframes (and different amounts). Interestingly, the pattern displayed seems randomized but again may not tell us much about user engagement, however it may tell us something else..



```

plt.bar(df['session_projects_added'], df['True_duration'])
plt.xlabel('session_projects_added')
plt.ylabel('True_duration')
plt.show()

plt.scatter(df['session_projects_added'], df['True_duration'])
plt.xlabel('session_projects_added')
plt.ylabel('True_duration')
plt.show()
```

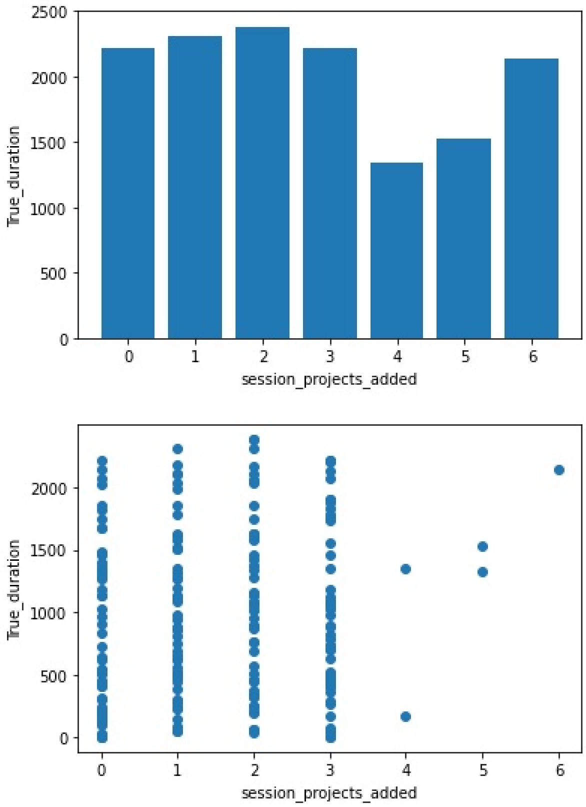
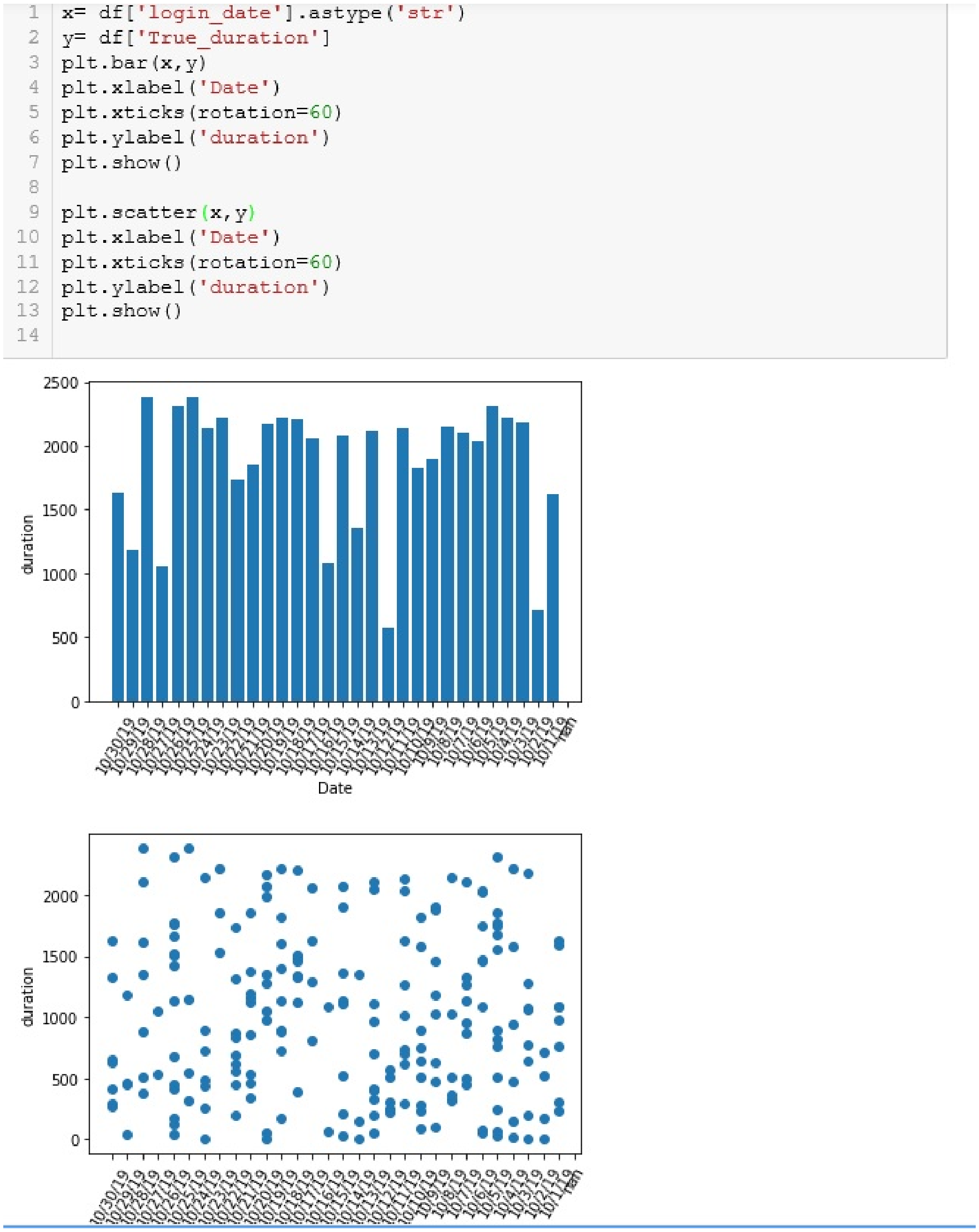


Figure 5: Figure of Session project vs True duration in both scatter and Bar graphs

This figure of projects tells us that customers will typically post projects from ranges of 0 to 3. It also tells us very few will post more than that and are on slightly less often than those at the higher end spectrum. This can potentially be used to extract infromation on user engagement, but still does not seem sufficient.

Last figure, shown below on the next following page, is about the login date vs True duration. The purpose is to observe how often customers are on and for how long. The idea of the graph is to use for future works, to see which days (weekends, weekdays, holidays) are users often on and interacting. Otherwise for now, its purpose is to analyze the peaks and dips of the customers and observe what do customers during these peaks and dips like to do.



```
x= df['login_date'].astype('str')
y= df['True_duration']
plt.bar(x,y)
plt.xlabel('Date')
plt.xticks(rotation=60)
plt.ylabel('duration')
plt.show()

plt.scatter(x,y)
plt.xlabel('Date')
plt.xticks(rotation=60)
plt.ylabel('duration')
plt.show()
```

Figure 6: Figure of Bar and Scatter for login date vs True duration

## 2.2 Analyzing the peaks and dips

I will be only picking one peak and one dip to analyze, the peak date(highest point) I choose is 10/28/19 and the dip date (lowest point) is 10/12/19. The code for picking these dates will be similar to the ones posted above but the difference is specifying the dates, so I will provide figures of each categories below. To get either high or low days, I generate a new dataframe (again using pandas) and state to use the specific dates I provided above.

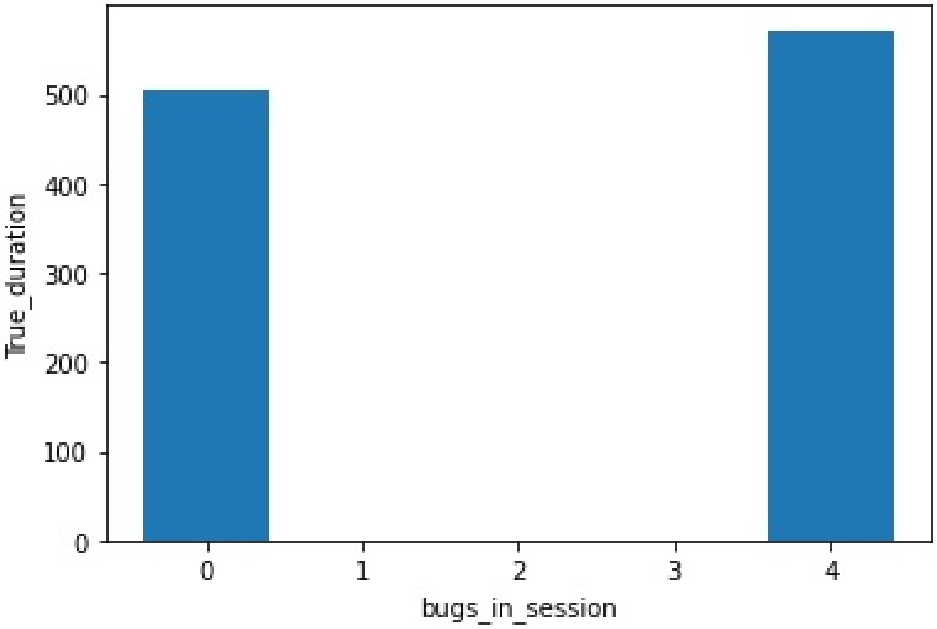


Figure 7: Figure of Bar graph of number of bugs vs True duration (Low day)

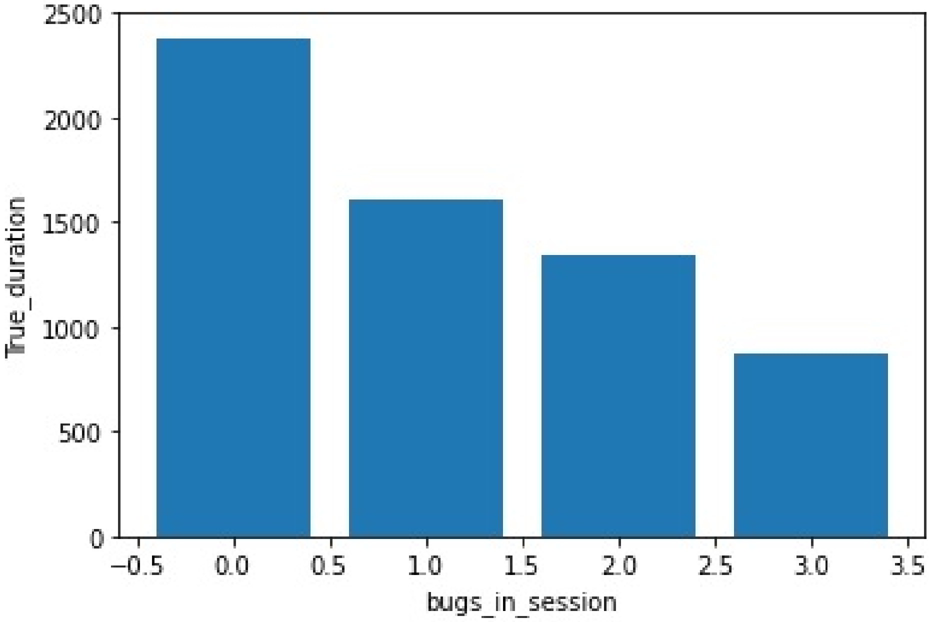


Figure 8: Figure of Bar graph of number of bugs vs True duration (high day)

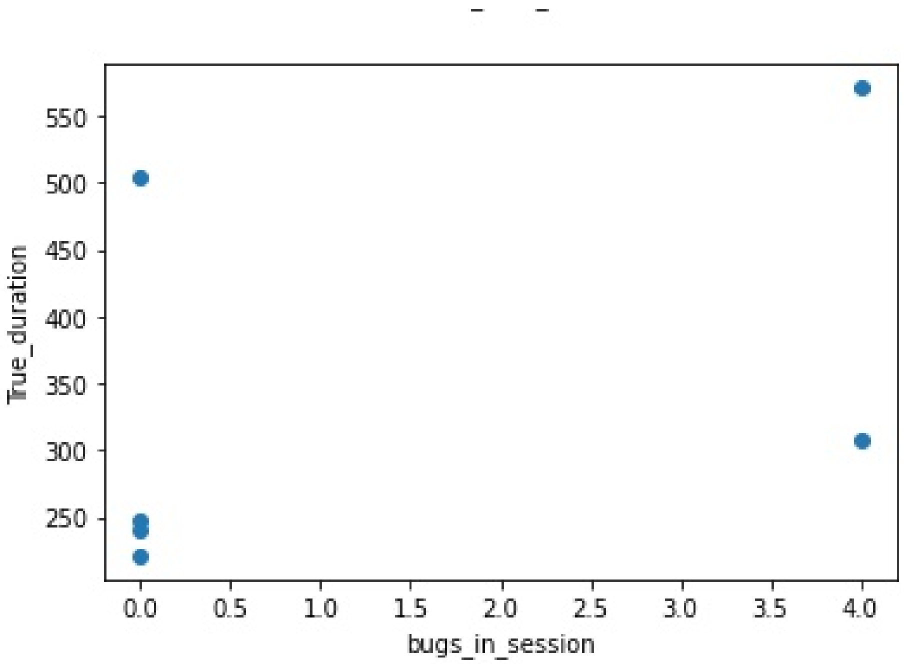


Figure 9: Figure of Scatter for number of bugs vs True duration(Low day)

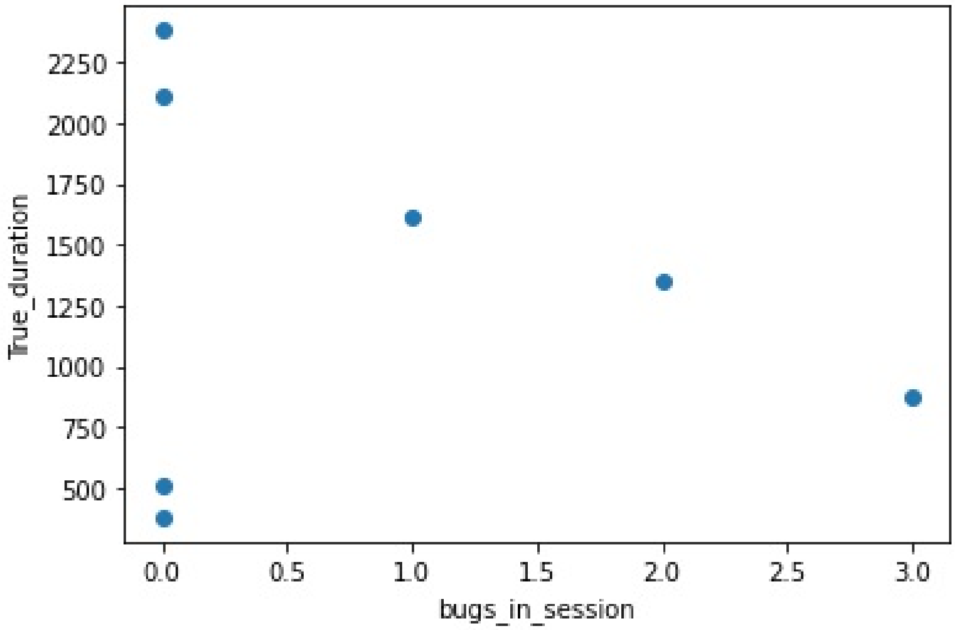


Figure 10: Figure of Scatter graph of number of bugs vs True duration (high day)

From the 4 figures above produced when setting bugs vs True duration, we can see that during low days the number of bugs experienced will be on the extreme ends ( either 0 or 4 when looking on the scatter graph specifically). While comparing to the high days, you would see a more evenly distributed number of bugs (Scattering graph shows this case).What this information provides is regardless of the days, customers will generally experienced bugs and the frequency of seeing them may be dependent on the day or activity. We can use the bugs feature to see what sort of activity customers tend to do when bugs are not prevalent.

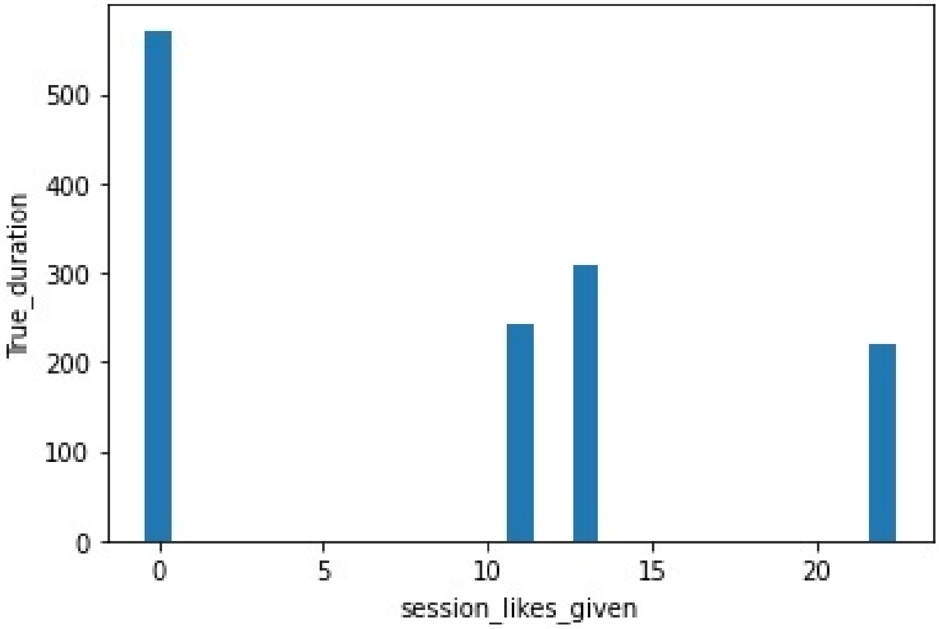


Figure 11: Figure of Bar for likes vs True duration(Low day)

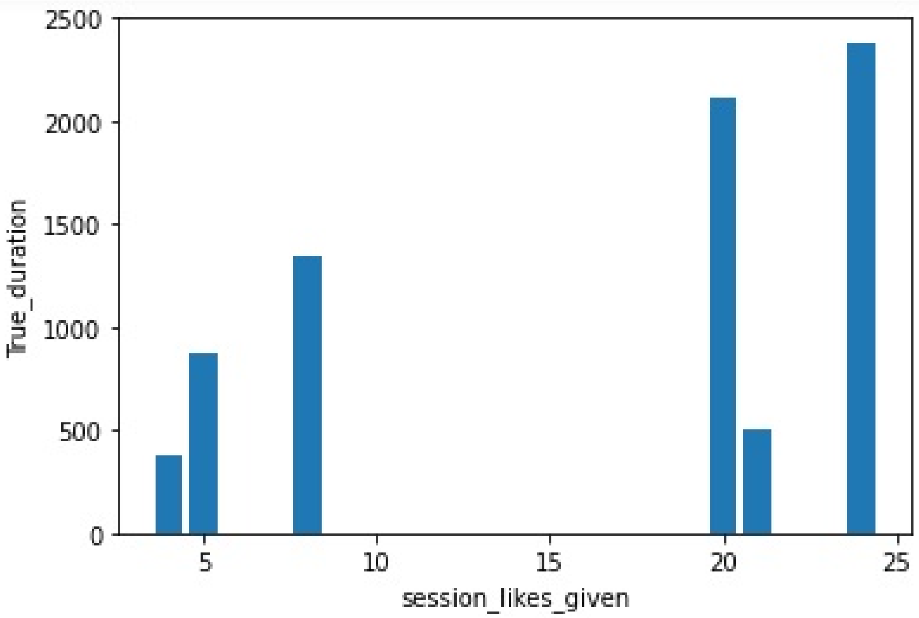


Figure 12: Figure of Bar graph of number of like vs True duration (high day)

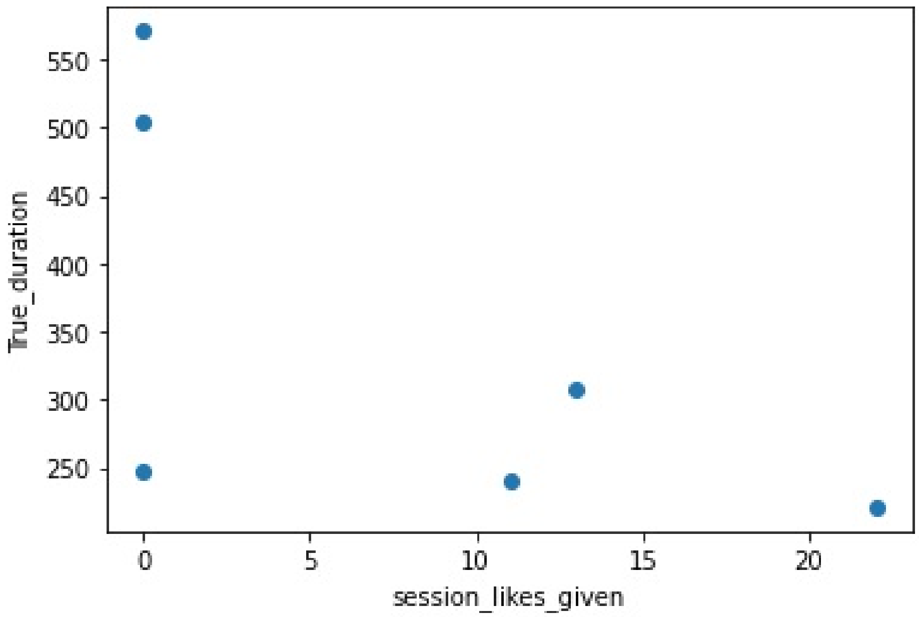


Figure 13: Figure of Scatter for likes vs True duration (Low day)

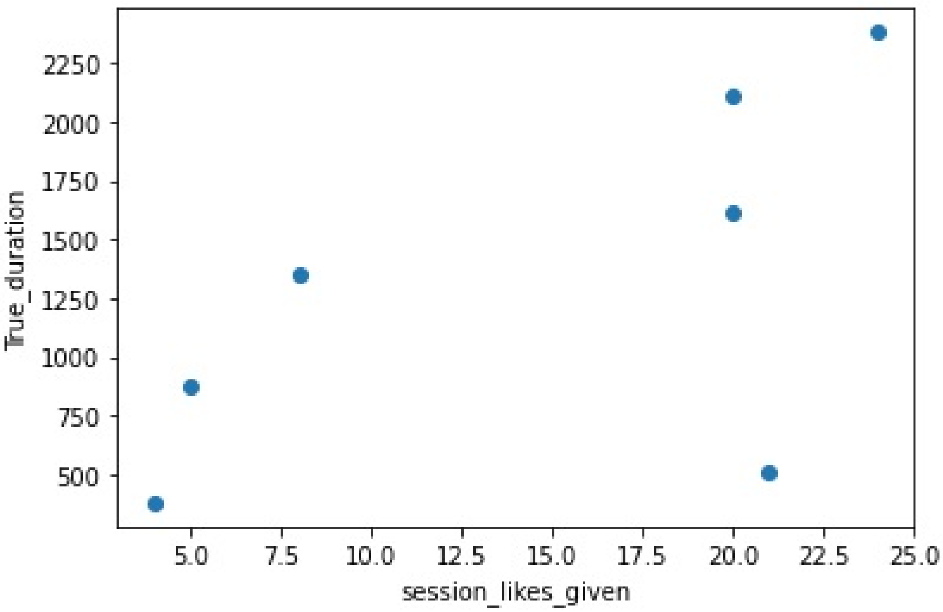


Figure 14: Figure of Scatter graph of number of likes vs True duration (high day)

From these 4 figures, we notice that people still generally try to input likes with the only notable difference is how long the individual spends their time on Showcase's product. If the number of likes is a variable for user engagement, we would expect to see even in the low days that customers would spend more time sending likes. But this is evidently not the case looking at the low day graphs. The number of likes may provide different information about our customer, and I will try to explain more about this in the next section.

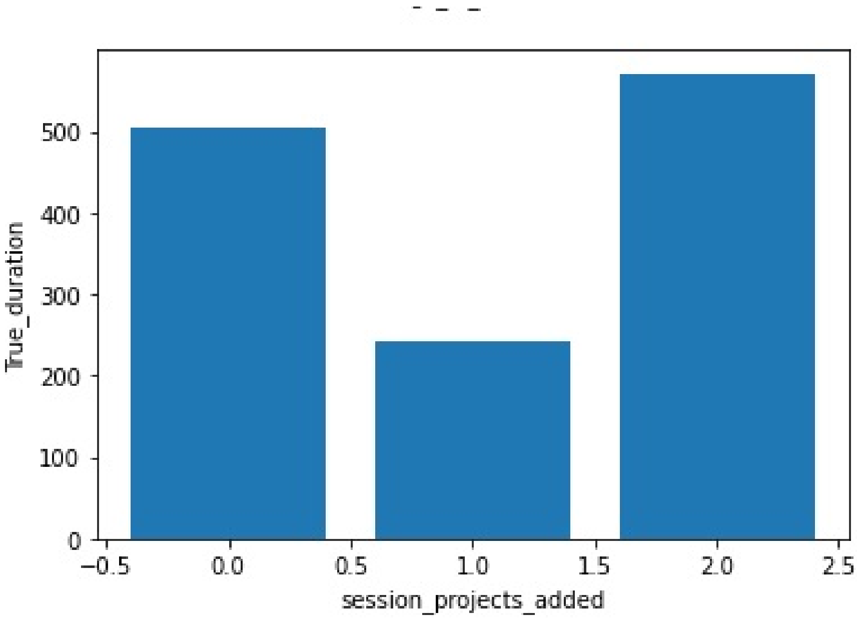


Figure 15: Figure of Bar for project vs True duration (Low day)

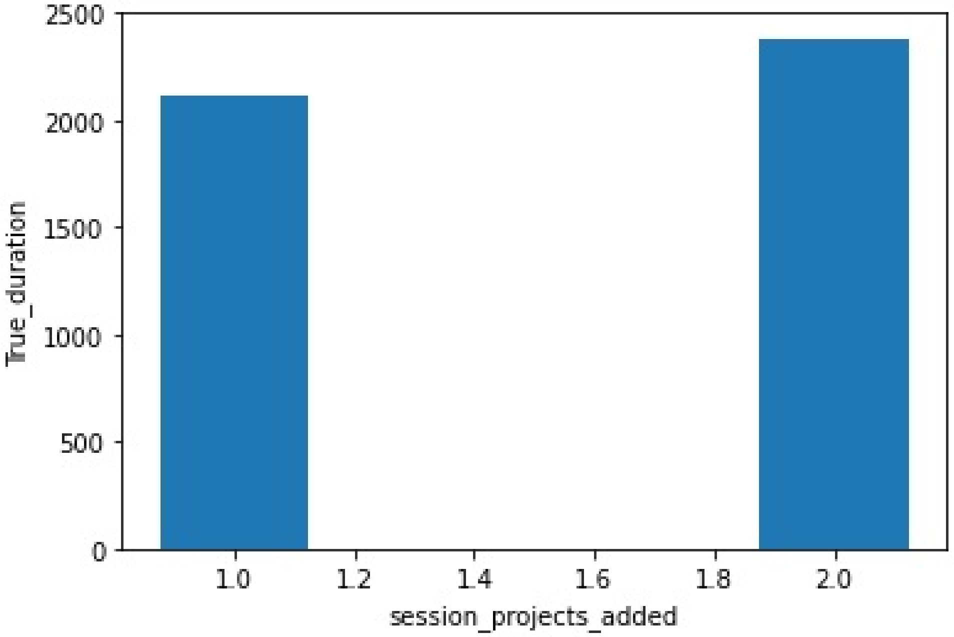


Figure 16: Figure of Bar graph of number of projectvs True duration (high day)

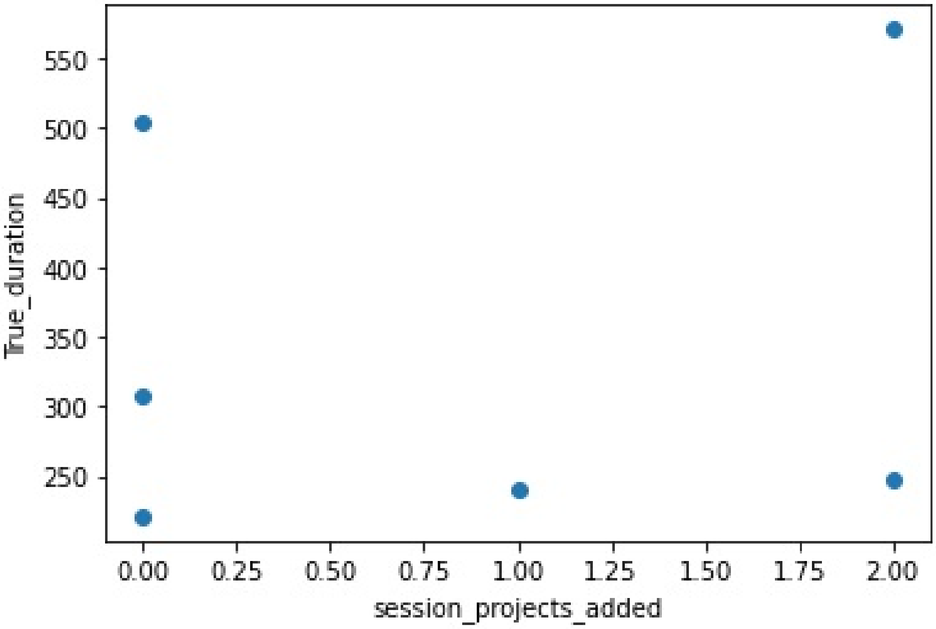


Figure 17: Figure of Scatter for project vs True duration (Low day)

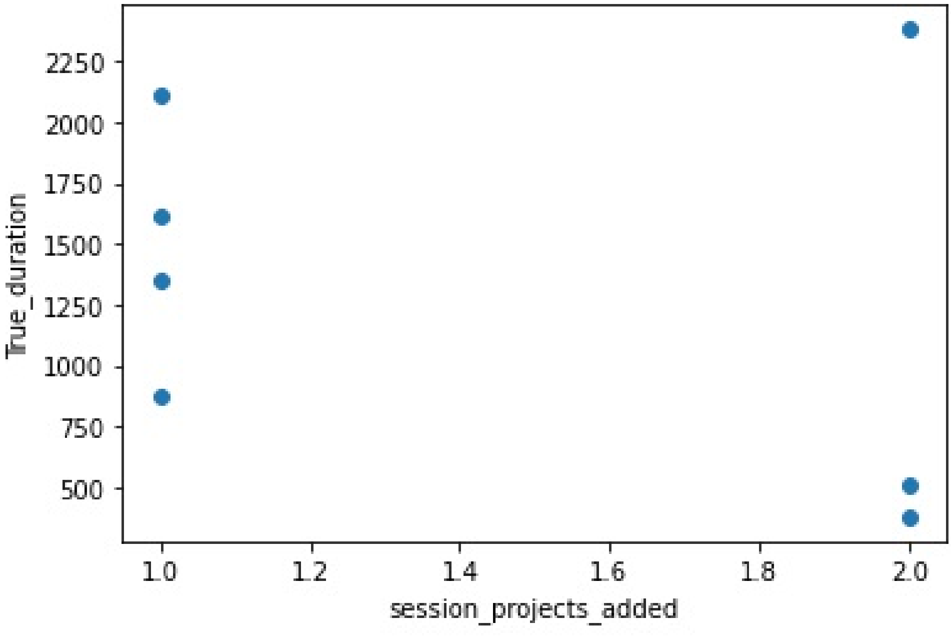


Figure 18: Figure of Scatter graph of number of project vs True duration (high day)

Interesting the number of projects portray a different picture. During the high days, customers will post (at minimum) one project and maximum two while during the low days, customers will do at a similar trend (with the minimum number of projects being 0). During the low day, customers will still tend to spend time posting projects. In the next section, I will see if the number of projects is a sign of user engagement.

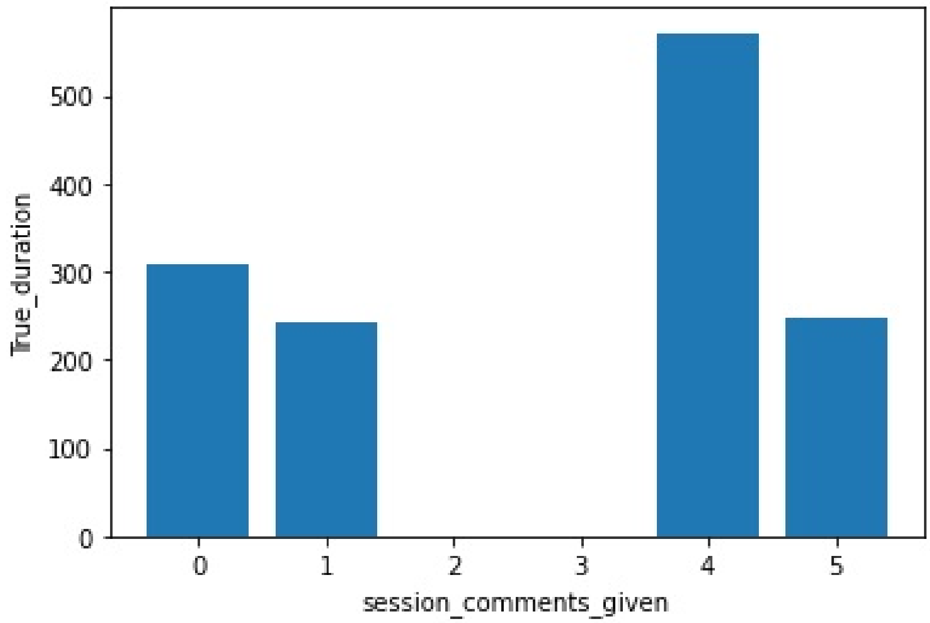


Figure 19: Figure of Bar comment vs True duration (Low day)

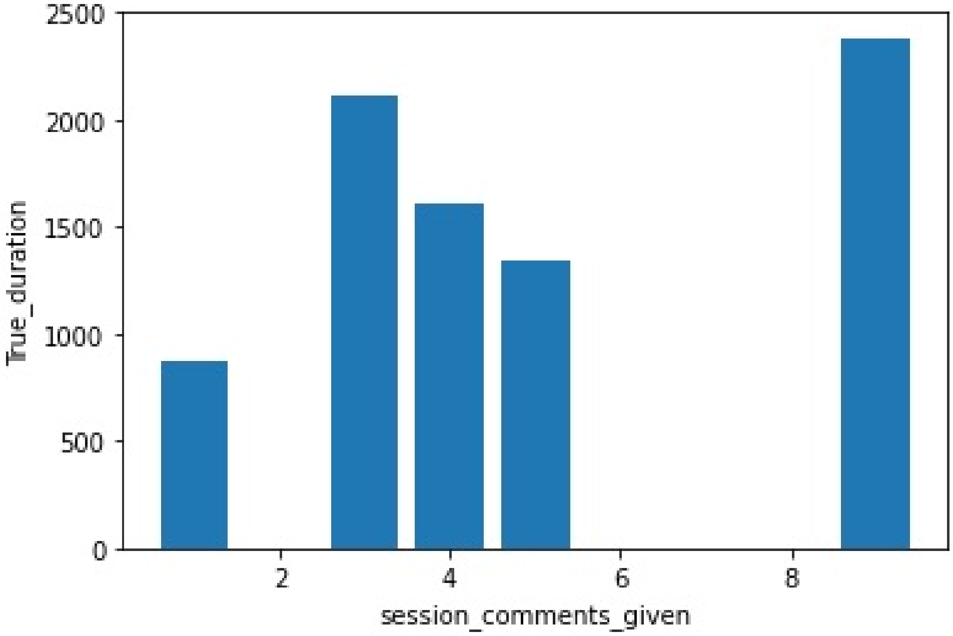


Figure 20: Figure of Bar graph of number of comment vs True duration (high day)

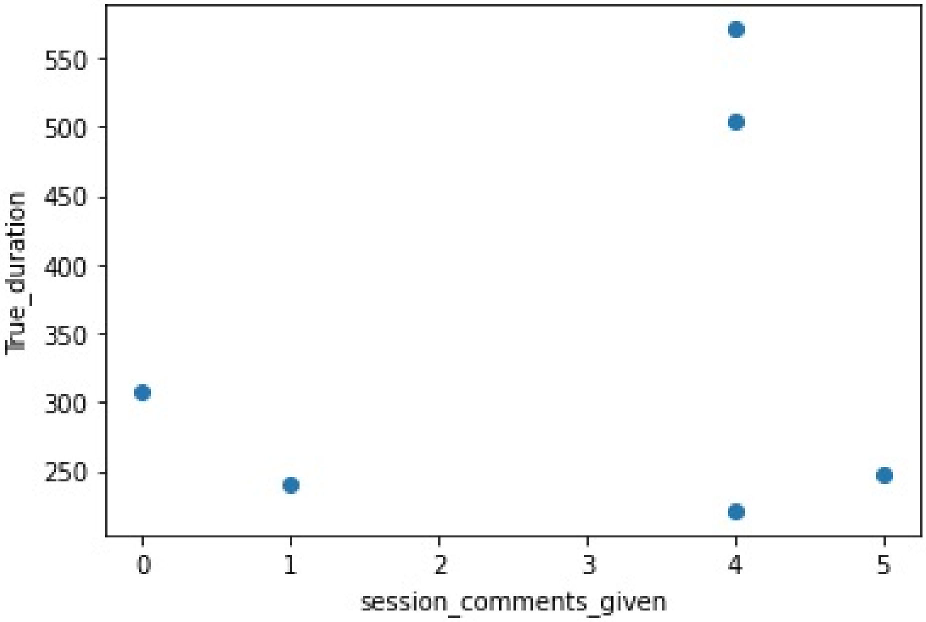


Figure 21: Figure of Scatter for comment vs True duration (Low day)

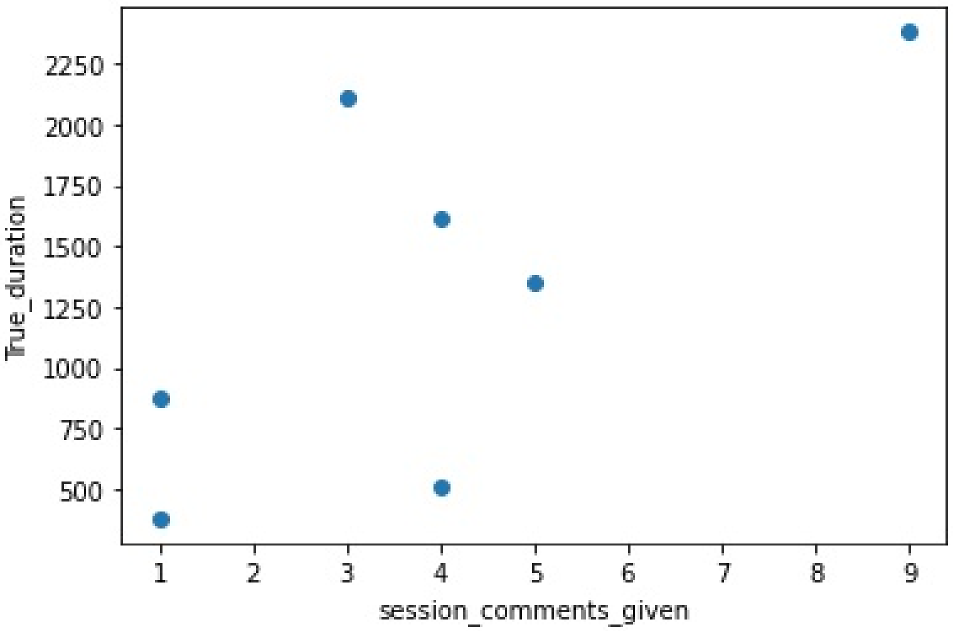


Figure 22: Figure of Scatter graph of number of comment vs True duration (high day)

These 4 figures shows us how much time consumers spend on commenting during both days. The high day scatter and bar indictate to us that people tend to comment more and spend more time doing so compared to the low days. There are some points where customers post for longer durations on low days, however its possible they may be outliers. This information may be useful when we try to isolate the data and see if comments is a good indicator of user engagement.

## 2.3 Isolating data

In this area, we will try to create dataframes that isolate the variables to see if they have any effect on the other sections. For example, we will generate a dataframe with data containing no bugs and see if they number of projects increase or decrease and if its frequency increases or decreases.

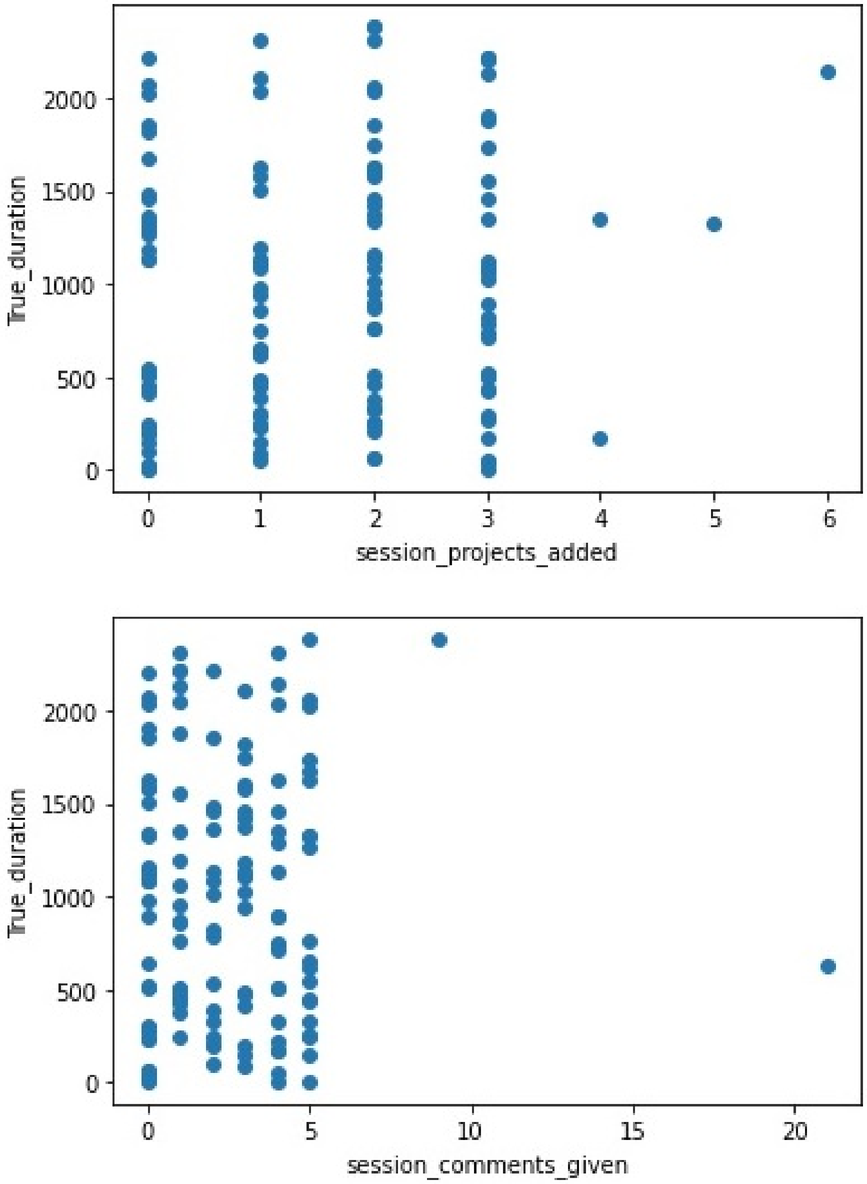


Figure 23: Figure Scatter of projects and comments when looking at data with no bugs)

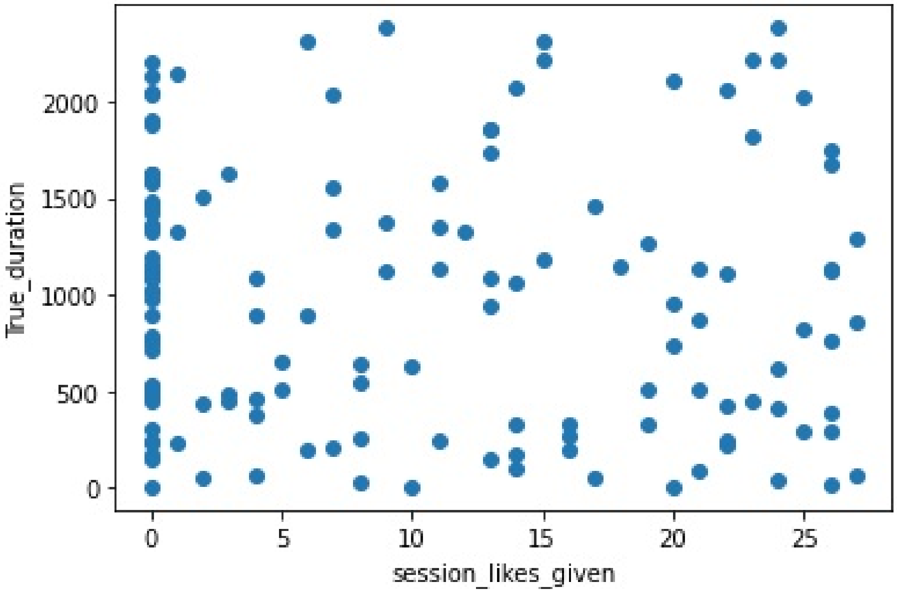
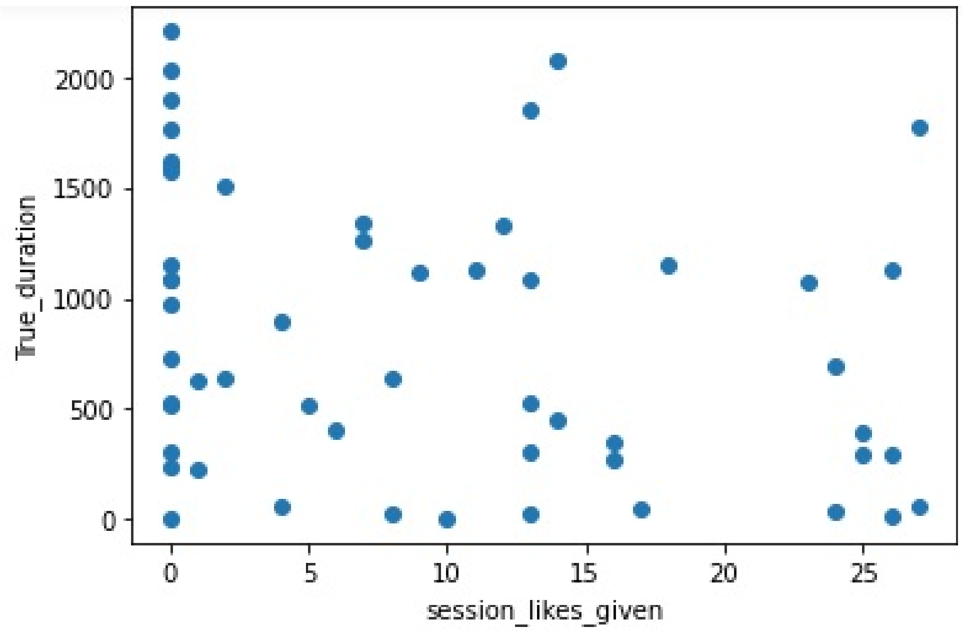
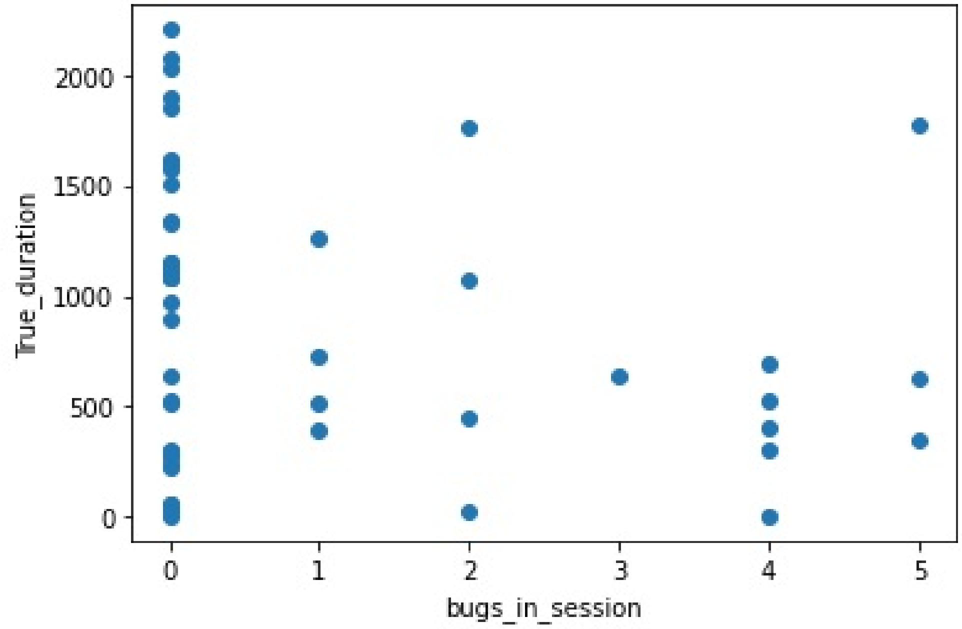
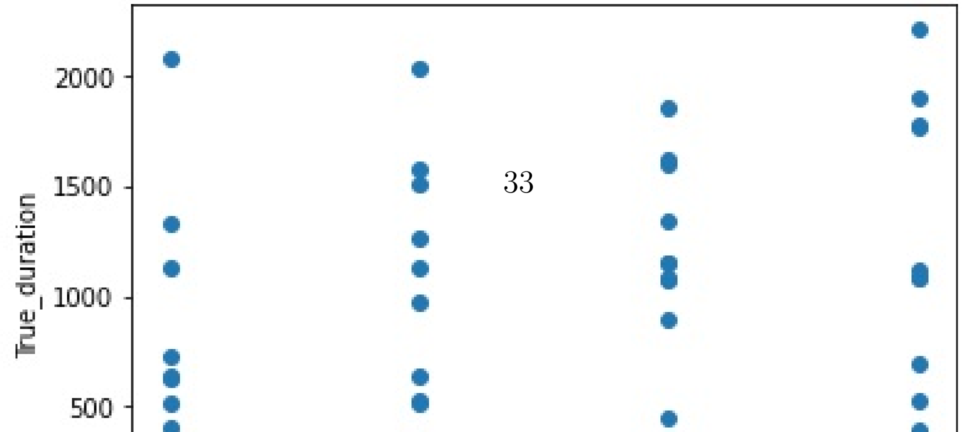


Figure 24: Figure Scatter of likes when looking at data with no bugs)

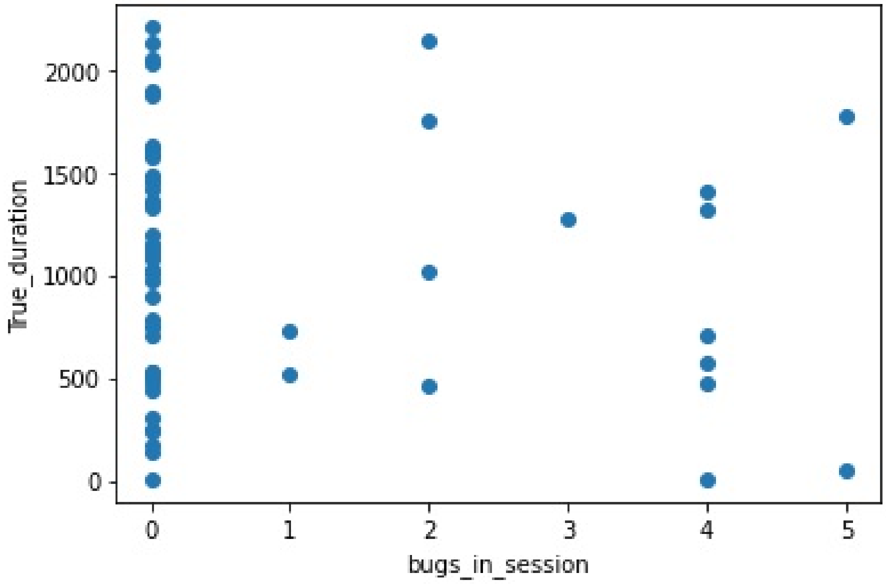
For the two figures above, we isolate for data points that has no bugs and observe the trends that other features portray. What we found is the number of comments and projects occur in a greater frequency (based on the color or the number of dots in the graphs).This is a useful trait to tell us that without bugs, people are more likely to engage in commenting or putting in projects. The number of likes appear to be random and its ”intensity” does not generate a definite pattern.

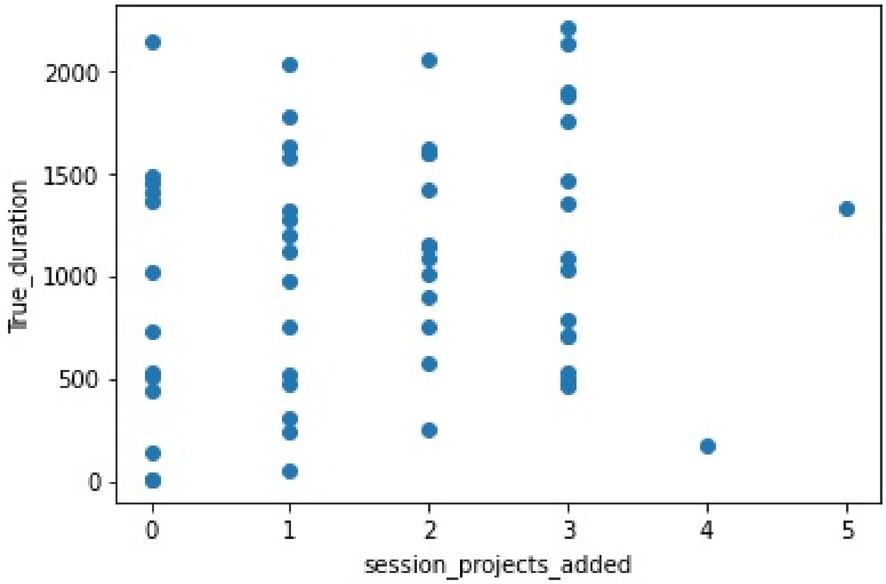

True_duration
0
500
1000
1500
2000
0
5
10
15
20
25
session_likes_given

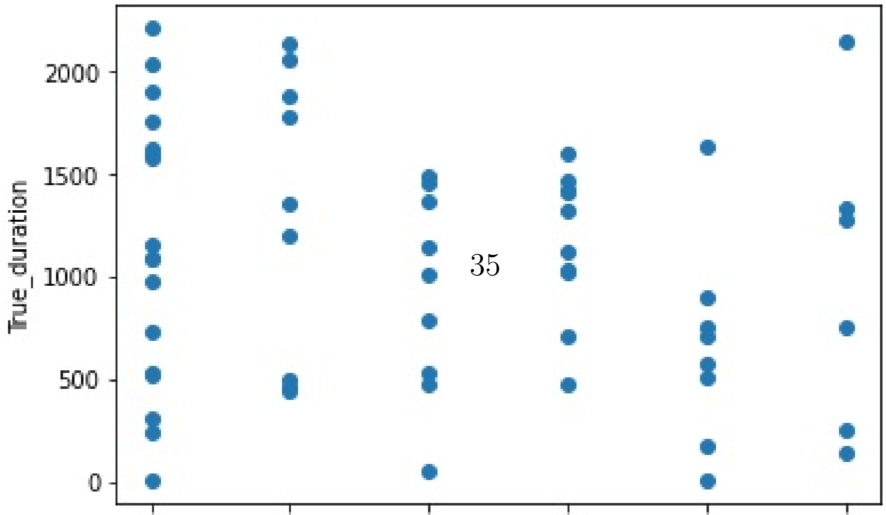

True_duration
0
500
1000
1500
2000
0
1
2
3
4
5
bugs_in_session


True_duration
500
1000
1500
2000

When looking at data that does not comment, the intensity of projects seems to decrease (the number of dots appearing seems to decrease). The number of bugs seems to vary and interestingly enough, the number of likes appear to take the same form as the data with no bugs.The number of likes may be constant for the rest of the data types and probably will not tell us anything about user engagement.


True_duration
0
500
1000
1500
2000
0
1
2
3
4
5
bugs_in_session


True_duration
0
500
1000
1500
2000
0
1
2
3
4
5
session_projects_added


True_duration
0
500
1000
1500
2000

For data with no likes, customers seem to put projects and comments in the same frequency as the dataset without bugs. And bugs seem to also be prevalent with customers who do not use the like feature. This tells us that while the like feature does seem to be used a lot by customers, it does not affect any of the other variables present and it is not a form of user engagement.

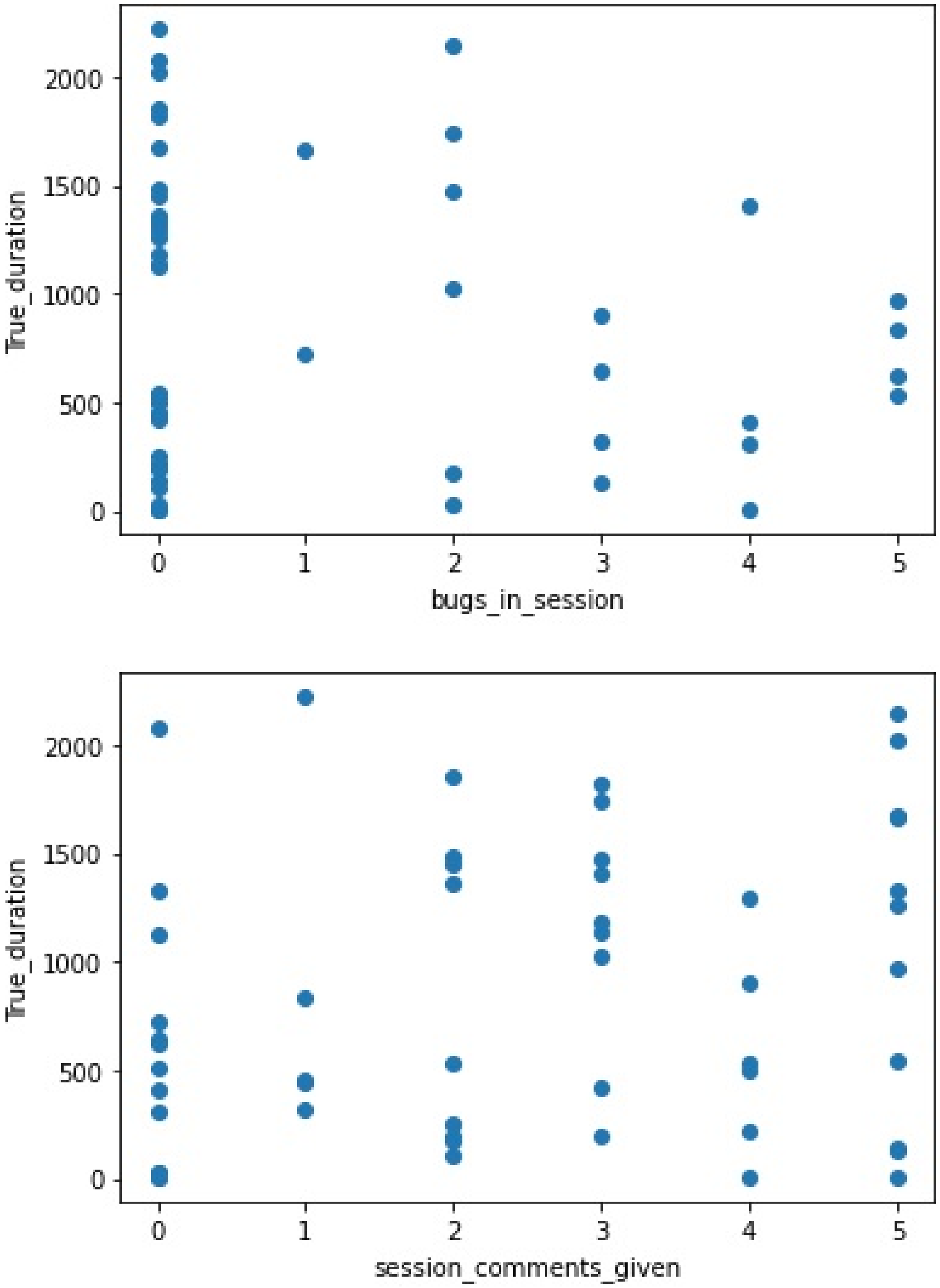


Figure 27: Figure Scatter of comments and bugs with data that has no projects)

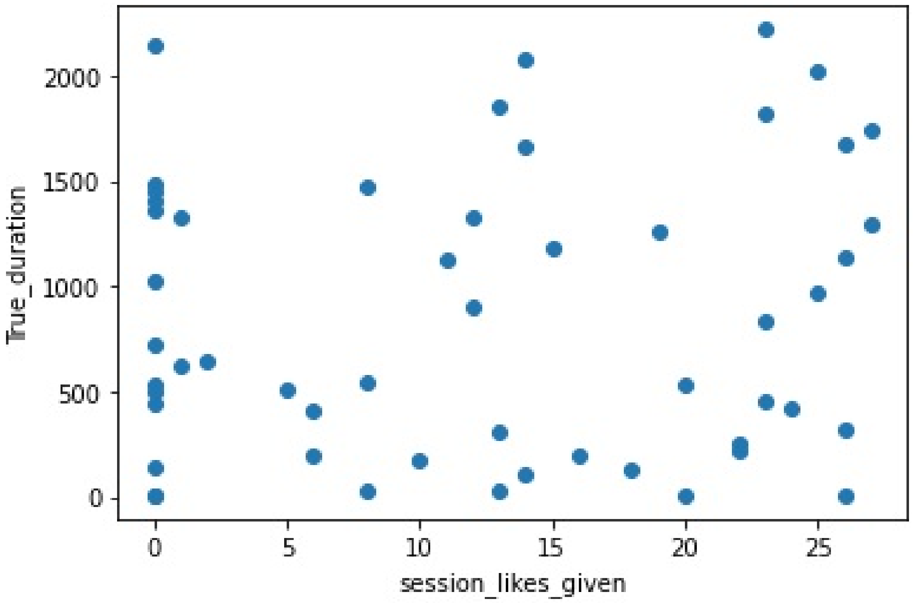


Figure 28: Figure Scatter of likes with data with no projects

For data without projects, comments seem to be affected and its frequency tends to be less. The number of likes and bugs seems consistent.

## 2.4 Conclusions and Future Works

User engagement is defined as the interaction between the product and its customers and can be found by using two variables: number of projects and number of comments. The number of bugs tells us how customers are less willing to engage in product if there are bugs in it, you expect customers who do not encounter bugs tend to use more of Showcase's features. The number of likes is an interesting feature that does not tell us user engagement but rather it is used for a different purpose. It is possible the reason why the number of likes is consistent regardless of the data you remove ( but users would spend less time on the feature during low days) is because the like feature is something people use to push other's projects into view. If we want the like feature to become a user engagement, we should find ways to get users to spend more time on the topic they are about to like and get them to interact with it. I believe for other future applications we can apply Unsupervised Learning from Machine Learning to see what sort of features would an algorithmn use to determine whether or not a certain dataset has user engagement.