**Documentation of Social Media Vs. Emotional Well-Being Project**

* *Ask & Prepare Phase*

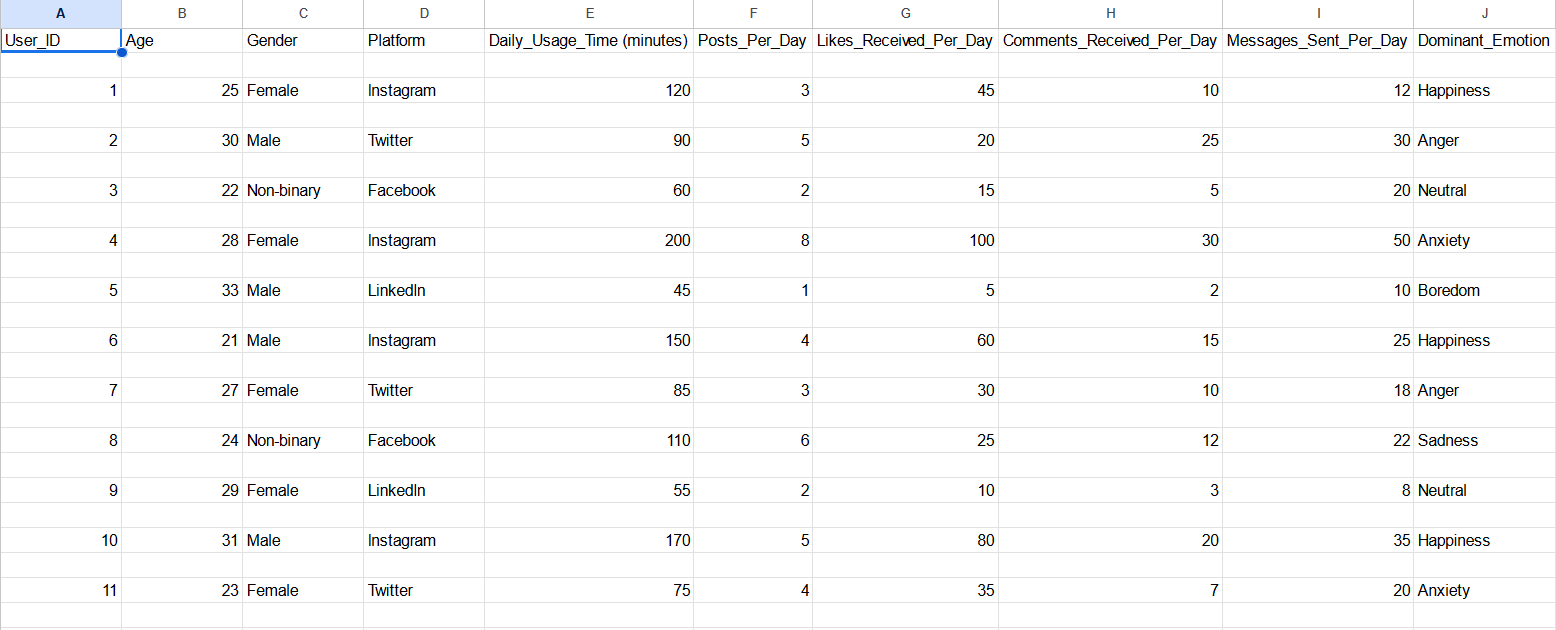
In this phase, being my first project, I first searched for an interesting and relevant database. I stumbled upon this one “<https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being?select=val.csv>” on the popular website Kaggle. After I found it, I decided to imagine a business scenario where I might use this data.

I agreed on the following scenario, I am a data analyst working for an upcoming app that is supposed to improve your wellbeing by restricting your access to certain apps that may distract or provide emotional distress to its users.

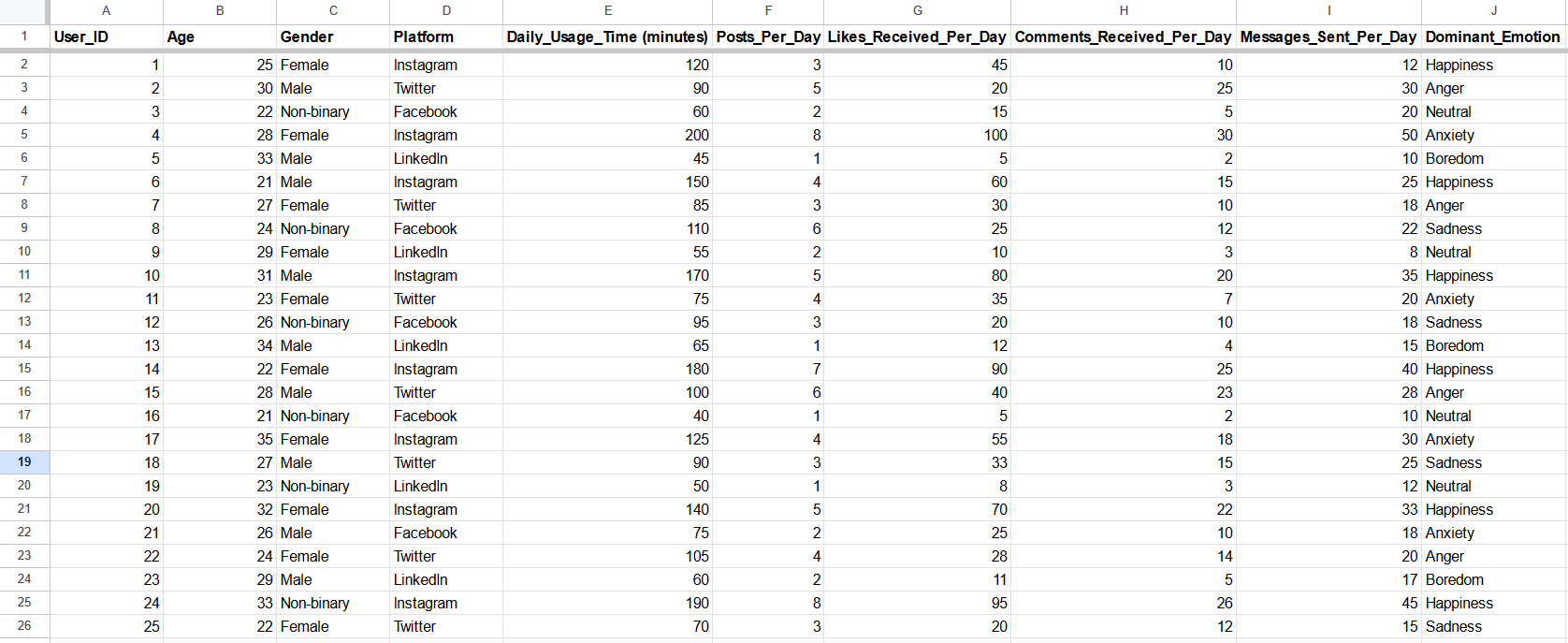
The data was collected through a hypothetical survey where participants were asked about their daily social media usage, including the amount of time spent on different platforms, the number of posts made, likes and comments received, and messages sent. Additionally, participants reported their dominant emotional state at the end of each day.

* *Data Cleaning*

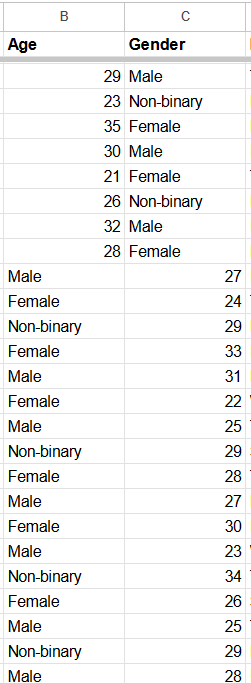
**Step 1**

Once I completed the ask & prepare phase, I opened the dataset with Google sheets, so that I could understand how everything is laid out and formatted.  
  
The first problem is the datasets' layout:  


As you can see, there are too many spaces between the rows, also the headers are not recognized as headers, so we need to fix these 2 problems.  
  
  
After I created the header row, by sorting the data, I deleted 1000 empty rows from the dataset:



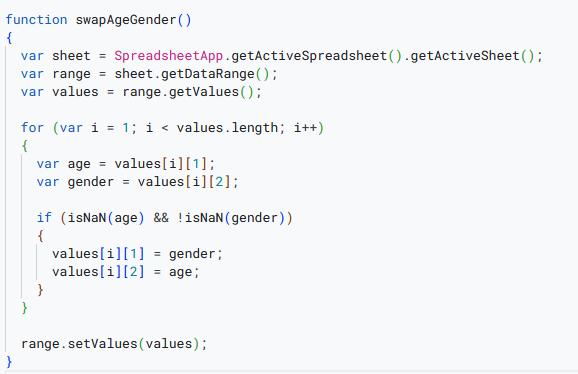
**Step 2**



The next problem I spotted was that some rows had the gender and age swapped:

This is just one example, there were around 120 rows with this issue.

Using a simple C++ script from the “app script” extension in google sheets, I swapped all the wrong values from the dataset. Here is the script:



**Step 3**

Now, using filters I checked for any outliers, any values that shouldn’t be possible/are wrong. Luckily, there weren’t any.

I also checked for data standardization, checked if all values were measured to the same scales, there were just a few mistakes in the daily\_usage\_time that I quickly corrected.

In the end, I checked for any NULL or missing values that might affect my analysis, now that everything is cleaned and nicely arranged, I can begin analyzing the data.

Link to Clean Dataset:

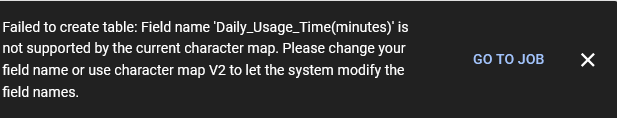
[SocialMedia\_Clean](https://docs.google.com/spreadsheets/d/1h70ZiIPuLLu304s0mthG5gYdIYtmbnfzwD8IbXR1f9U/edit?usp=sharing)

Link to Raw Dataset:

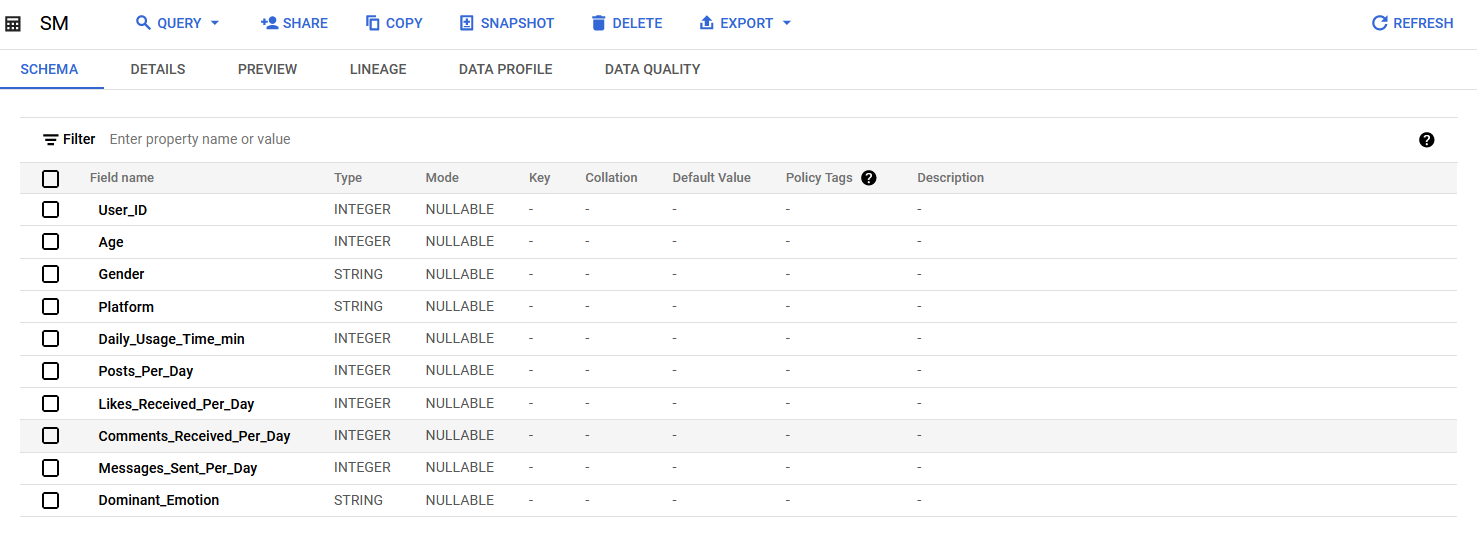
[SocialMedia\_Raw](https://docs.google.com/spreadsheets/d/1tAYxTPYfligxrPEUoX_vkDTXvbP7o3WQNqqFpPf4lNY/edit?usp=sharing)

* *Data Analysis*

I will be conducting the analysis using tools such as Google BigQuery(SQL), RStudio(R) and Tableau.

When trying to load the dataset, I get this error:

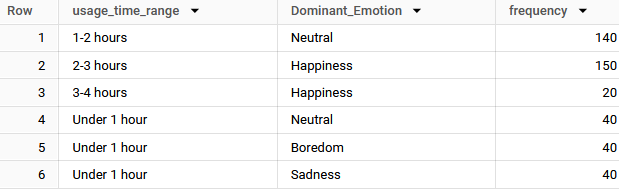
Most likely because the parentheses are not supported, let’s modify it as Daily\_Usage\_Time\_min and see if it works.

And just as I guessed, that was the problem, it works now. 

Taking a quick look, we can see that all the fields are formatted properly, so we won’t have to do any modifying on that part.

Starting with the SQL part, as there is only one table JOIN functions are not necessary, however I can still use SQL to find insights about the data.

I decided to test out a theory. I think that more daily usage will lead to more negative feelings. Let’s see if that’s true:



According to this dataset, it’s not!! That’s pretty surprising, this is the SQL code I used to arrive to this conclusion:

WITH time\_segments AS (

SELECT

CASE

WHEN Daily\_Usage\_Time\_min < 60 THEN 'Under 1 hour'

WHEN Daily\_Usage\_Time\_min >= 60 AND Daily\_Usage\_Time\_min < 120 THEN '1-2 hours'

WHEN Daily\_Usage\_Time\_min >= 120 AND Daily\_Usage\_Time\_min < 180 THEN '2-3 hours'

WHEN Daily\_Usage\_Time\_min >= 180 AND Daily\_Usage\_Time\_min < 240 THEN '3-4 hours'

WHEN Daily\_Usage\_Time\_min >= 240 AND Daily\_Usage\_Time\_min < 300 THEN '4-5 hours'

ELSE '5+ hours'

END AS usage\_time\_range,

Dominant\_Emotion

FROM `charming-mile-419712.SocialMedia.SM`

),

emotion\_counts AS (

SELECT

usage\_time\_range,

Dominant\_Emotion,

COUNT(\*) as frequency

FROM time\_segments

GROUP BY usage\_time\_range, Dominant\_Emotion

),

max\_emotion\_per\_segment AS (

SELECT

usage\_time\_range,

Dominant\_Emotion,

frequency,

RANK() OVER (PARTITION BY usage\_time\_range ORDER BY frequency DESC) as rank

FROM emotion\_counts

)

SELECT

usage\_time\_range,

Dominant\_Emotion,

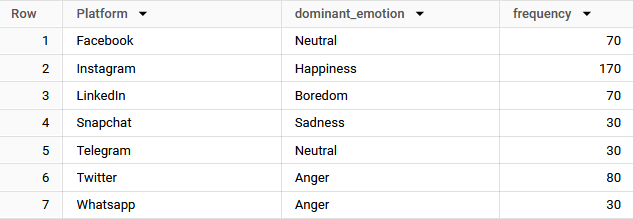
frequency

FROM max\_emotion\_per\_segment

WHERE rank = 1

ORDER BY usage\_time\_range;

Now, I’d also like to test out how each platform makes its users feel:



These are the results, from this insight we can flag these apps for the developers of the app, they could restrict their (the apps that have a negative dominant emotion) usage more strictly compared to the other apps:

The code I used to arrive to these conclusions was this:

WITH emotion\_counts AS (

SELECT

Platform,

dominant\_emotion,

COUNT(\*) as frequency

FROM `charming-mile-419712.SocialMedia.SM`

GROUP BY Platform, dominant\_emotion

),

max\_emotion\_per\_platform AS (

SELECT

Platform,

dominant\_emotion,

frequency,

RANK() OVER (PARTITION BY Platform ORDER BY frequency DESC) as rank

FROM emotion\_counts

)

SELECT

Platform,

dominant\_emotion,

frequency

FROM max\_emotion\_per\_platform

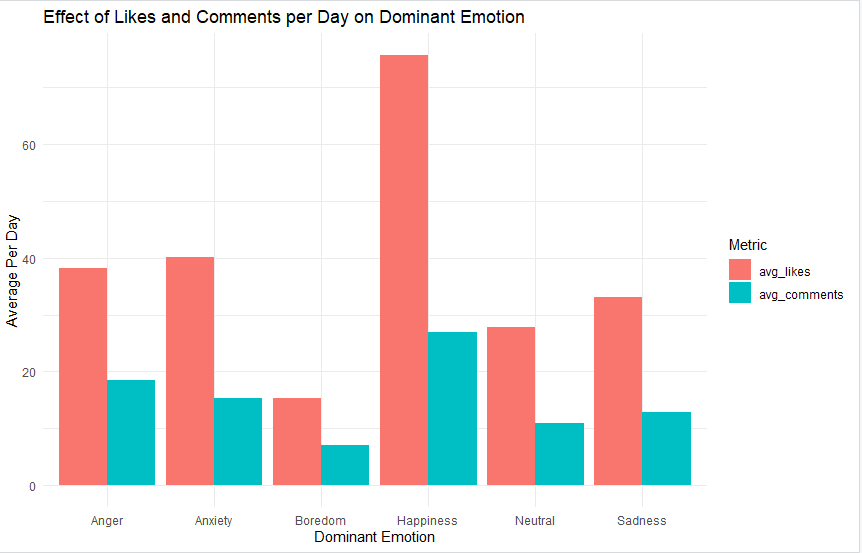
WHERE rank = 1

ORDER BY Platform;

I ended up running some more tests on this data with simpler MIN(),MAX() functions, but most of this data I already saw in the spreadsheets since it’s a relatively small dataset.

And now, I’ll switch to **R** to analyze some more data and create basic visualizations that will help once I switch to tableau.

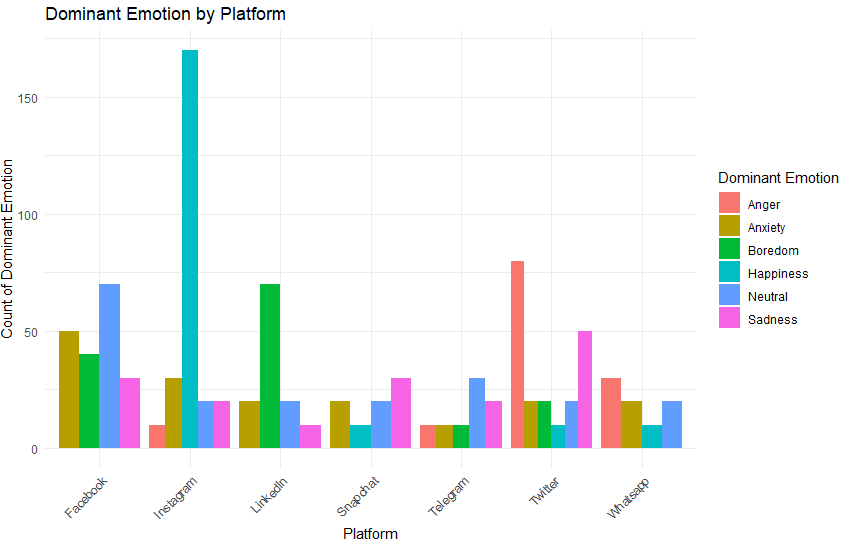
In R, I want to check how *likes and comments* affect the dominant emotion.

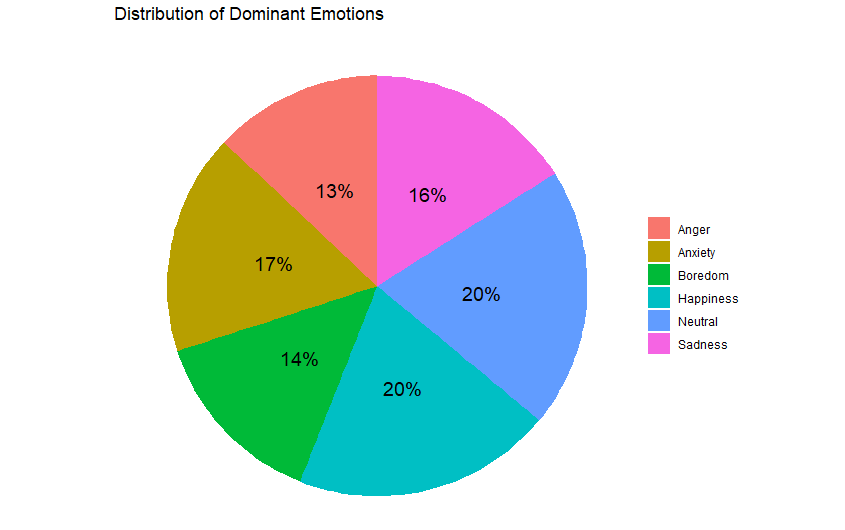


As you can see, happiness is closely tied with the amount of likes and comments. This metric can be very useful to the developers, allowing them to suggest the users to post more and engage more with the community. The code I used was this:



Now, I want to test 2 more visualizations in R, dominant emotion by platform and dominant emotion distribution:



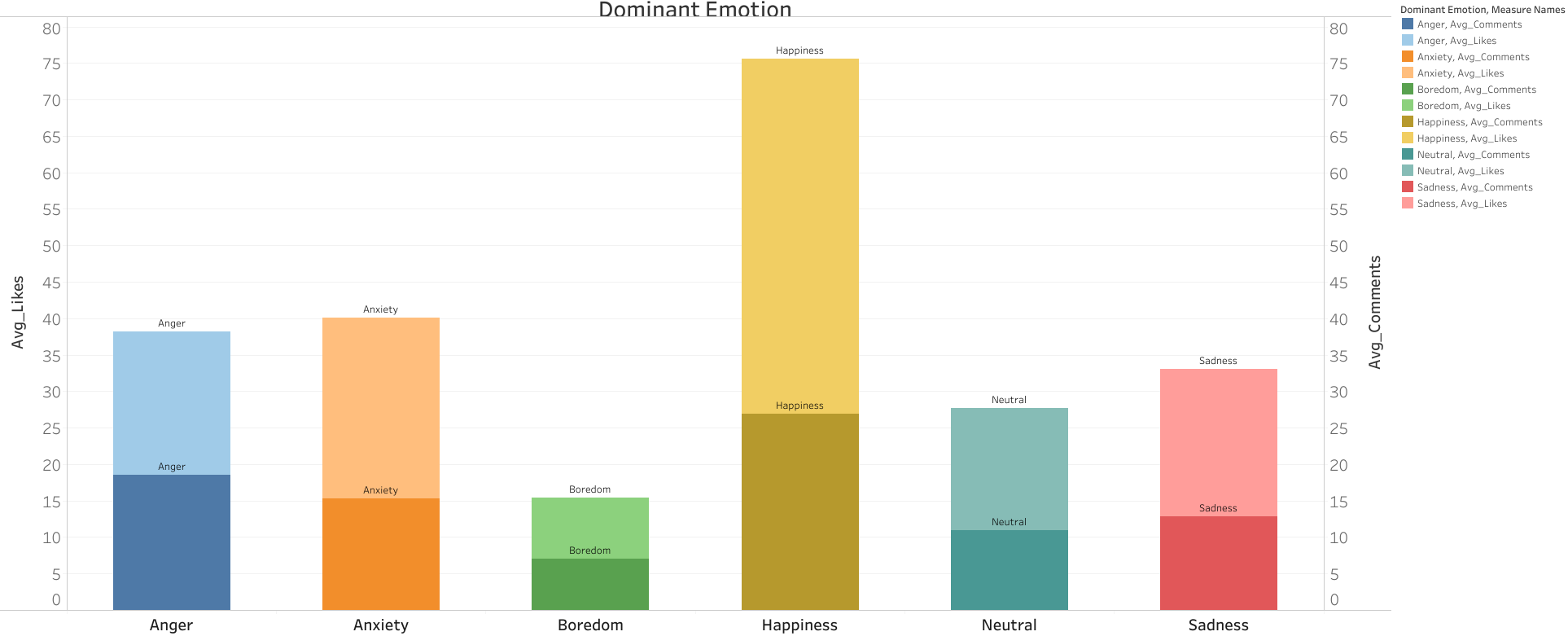


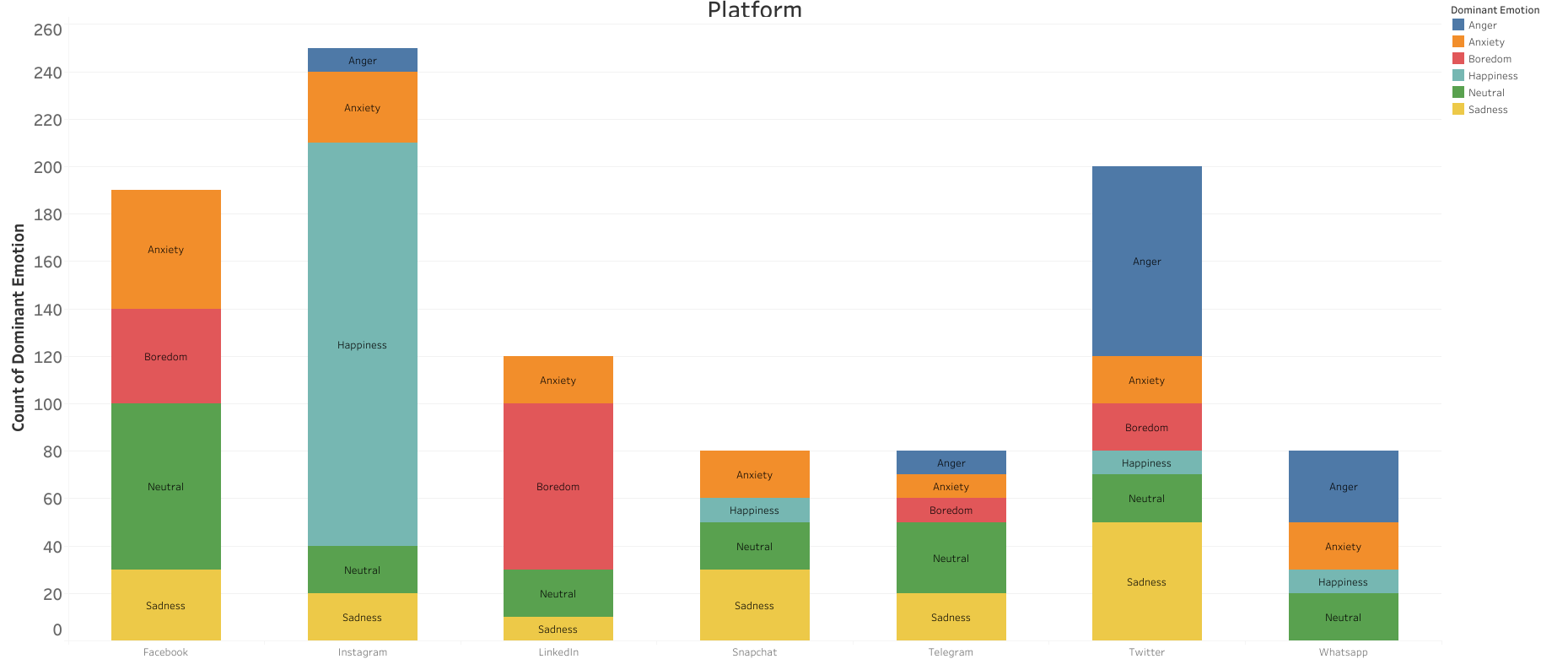
Now, we have confirmation for the data that SQL showed us as well, and I can get started on making proper charts and a presentation for the developers/stakeholders.

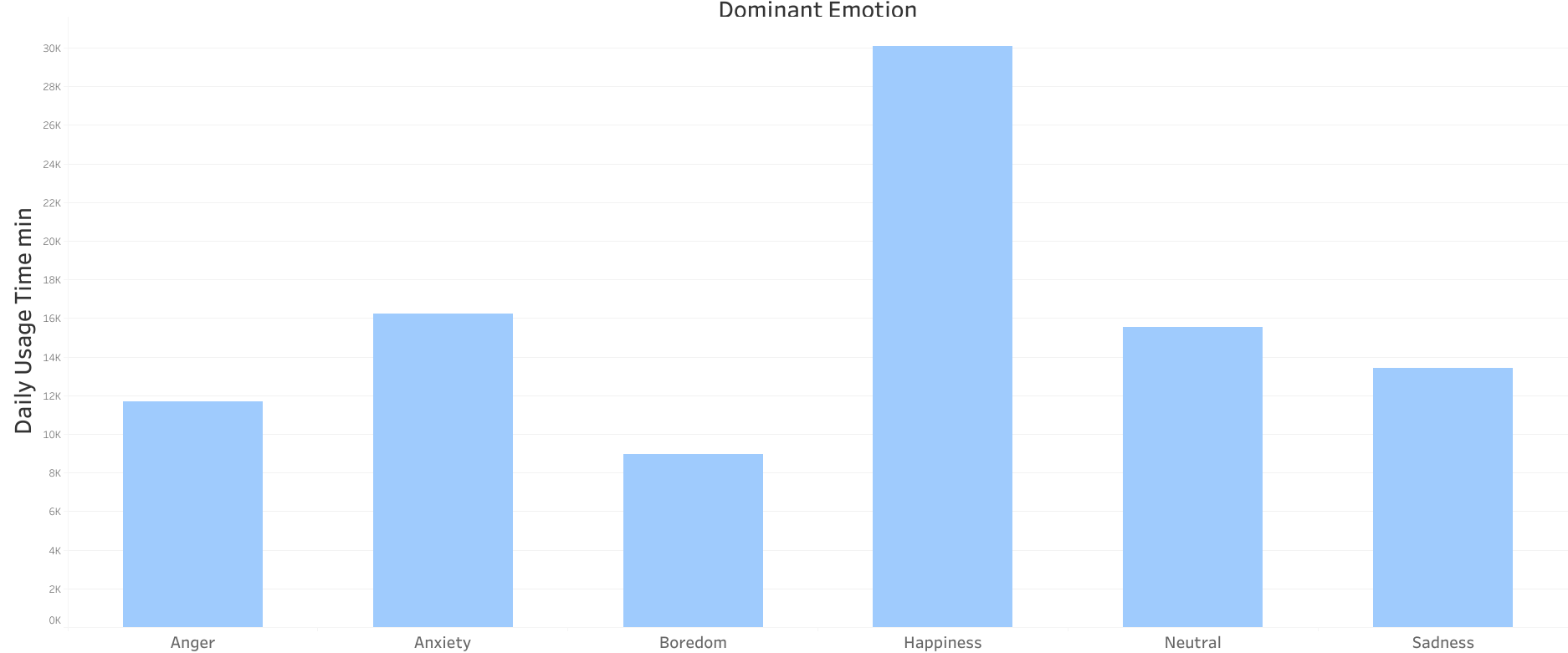
Oh, almost forgot the code:

https://drive.google.com/file/d/1yR5lsI55EIY1Q5\_VefLMeJw\_KdvgftvT/view?usp=sharing

Now, I switch to tableau in order to create better visualizations.







These are the visualizations I created for the final presentation, I didn’t have many problems until I created the first one, where I needed to use SQL functions in the Analysis>Create Calculated Field section of tableau, in order to create the proper visualization.