MI3-2 Establish Data to Analysis Plan

Group Name: SEM

Group Leader: Emily Feng (ejf9kwf)

Group Members: Mei Gilhousen (mlg8jc), Samarth Saxena (ss9nfb)

DS 4002, 11/9/2022

Hypothesis:

In the top 50 most used emoticons in the world, at least 50% of them denote more positive feelings than negative feelings, with a sentiment score above 0 on a scale from -1 to 1. Additionally, the correlation between the popularity of emojis (emoticons) and their sentiment scores are greater than 0.5.

Executive Summary:

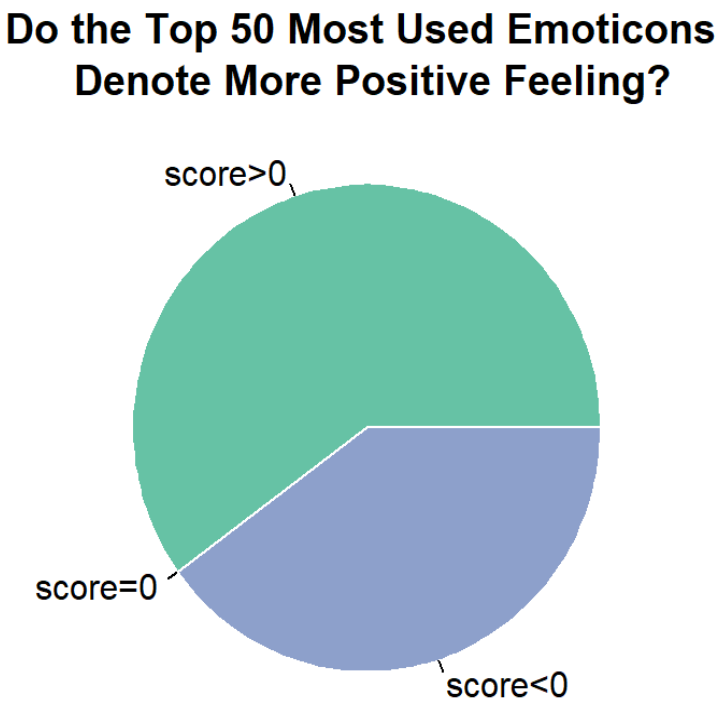
This document explores the data using the Kaggle emoji data set and the questions explored using the data including relevant plots. A summary of the data and hypothesis is included. Then an analysis plan of the to test the quantifiable hypothesis and related models are shown.

Data Discovery Findings:

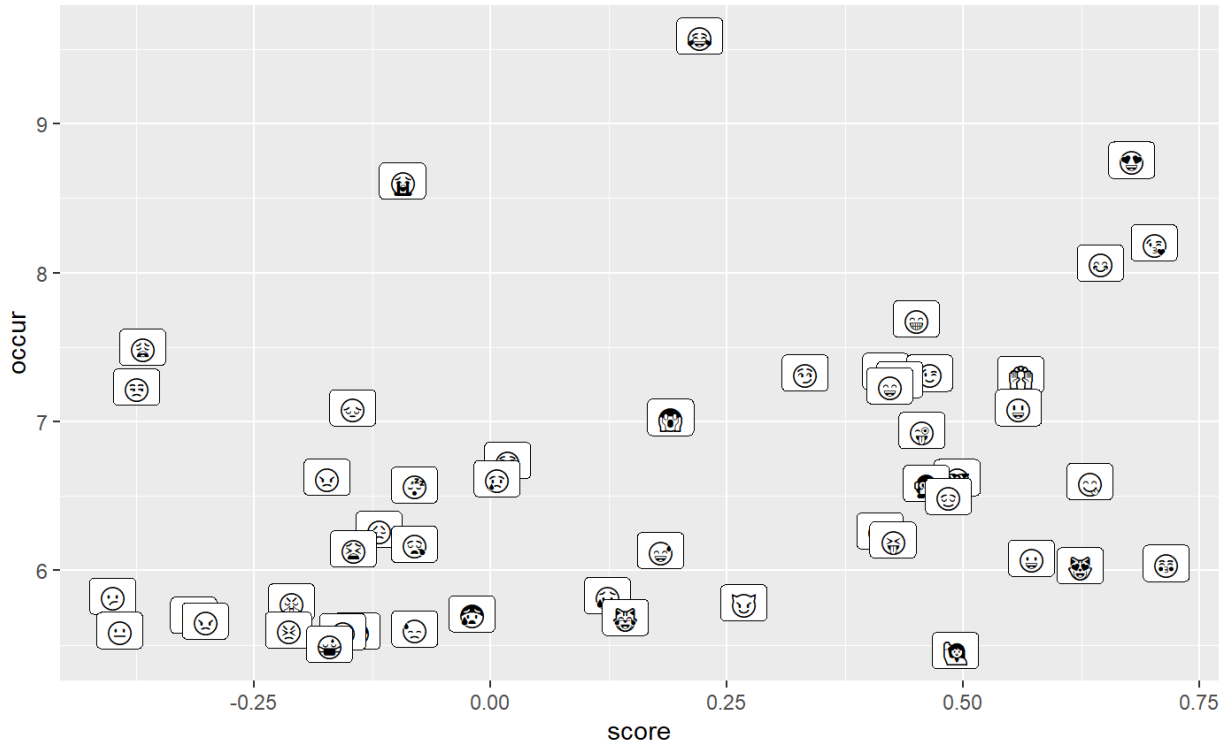
Link to dataset and dictionary: https://github.com/Lychee030/DS4002-M3

The data is retrieved from the Kaggle website as image data. There are 12 columns total, but our analysis focus is limited to the emoji observations, the occurrences, negative, neutral, positive and sentiment ratings. There are 752 observations total. Some questions we explored through the graphical analysis below were: What variables could be quantified numerically or categorically for analysis, what was the distribution of positive sentiment among the top 50 most used emojis, what was the score distribution by occurrence, and the linear regression for sentiment occurrence by sentiment scores. Some refinements to the hypothesis arose, as we had to make sure that there would be enough observations across the entire distribution that we analyzed (from -1 to 1 for positive sentiment). The dataset not only contains emoticons, but also includes pictographs, miscellaneous symbols, and others. Therefore, the first thing we did was to filter emojis based on their unicode.block. Then, we found 76 emoticons and limited them to the top 50 most used emoticons. Also, our goal is to find the correlation between occurrence of emoji and their sentiment scores. The number of occurrences could be really different. Thus, we analyzed log(occurrences) for our project instead. The revised dataset is shown below.

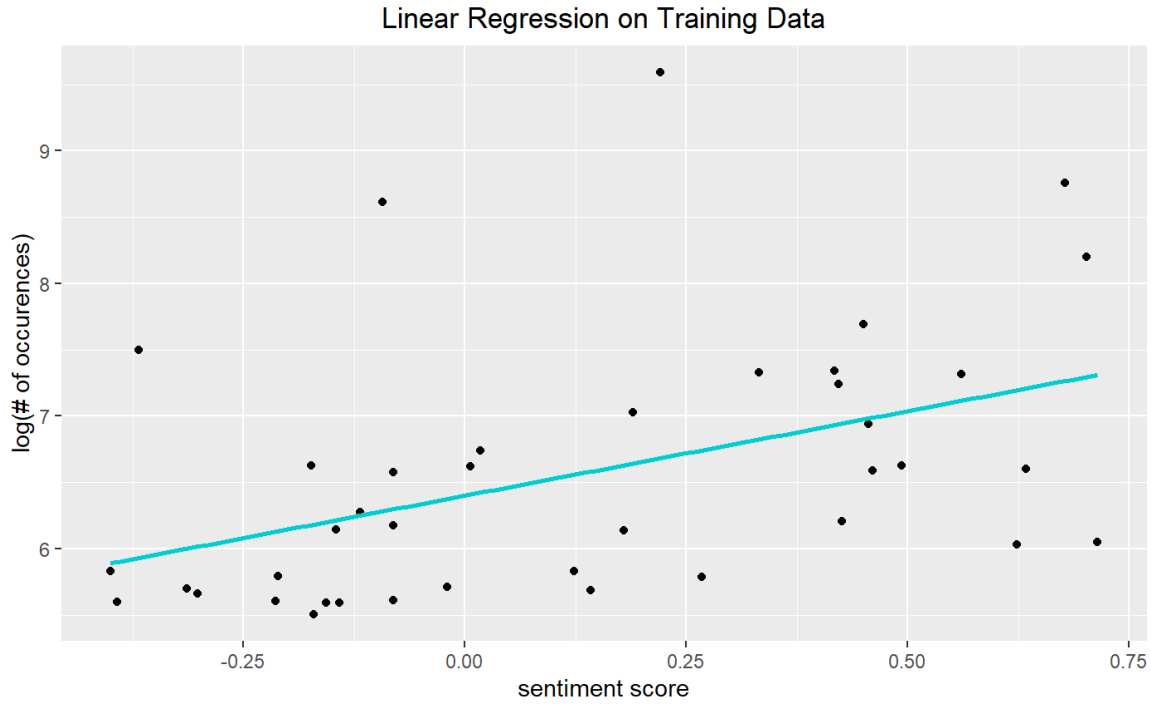
	Emoji	Unicode.codepoint	Occurrences	Position	Negative	Neutral	Positive	Unicode.name
1	😂	128514	14622	0.8051006	3614	4163	6845	FACE WITH TEARS OF JOY
4	😍	128525	6359	0.7652924	329	1390	4640	SMILING FACE WITH HEART-SHAPED EYES
5	😭	128557	5526	0.8033520	2412	1218	1896	LOUDLY CRYING FACE
6	😘	128536	3648	0.8544802	193	702	2753	FACE THROWING A KISS
7	😊	128522	3186	0.8133024	189	754	2243	SMILING FACE WITH SMILING EYES
11	😁	128513	2189	0.7961512	278	648	1263	GRINNING FACE WITH SMILING EYES
15	😩	128553	1808	0.8262140	1069	336	403	WEARY FACE
16	🙏	128591	1539	0.7938476	124	648	767	PERSON WITH FOLDED HANDS
18	😏	128527	1522	0.7649773	170	676	676	SMIRKING FACE
19	😉	128521	1521	0.8448327	151	513	857	WINKING FACE
20	🙌	128588	1506	0.7906000	152	358	996	PERSON RAISING BOTH HANDS IN CELEBRATION
21	🙈	128584	1456	0.7388809	238	350	868	SEE-NO-EVIL MONKEY
23	😄	128516	1398	0.7949726	191	426	781	SMILING FACE WITH OPEN MOUTH AND SMILING EYES
24	😒	128530	1385	0.8576206	819	266	300	UNAMUSED FACE
27	😃	128515	1206	0.7347822	86	361	759	SMILING FACE WITH OPEN MOUTH
28	😔	128532	1205	0.8661459	559	263	383	PENSIVE FACE
29	😱	128561	1130	0.7733132	298	319	513	FACE SCREAMING IN FEAR



To analyze whether or not at least 50% of the top 50 most used emoticons denote more positive feelings than negative feelings, we first made a pie chart. It is quite surprising to see that there are no emoticons with a sentiment score of 0. However, from the result, we can conclude that at least 50% of them have sentiment score greater than 0 on a scale from -1 to 1, which means that they are more commonly used accompanied with positive texts.



This graph shows the log(number of occurrences) of each emoji on the y-axis, with the sentiment score that each emoji received on the x-axis. This allows us to visualize which emojis portray positive sentiments and which ones portray negative sentiments.



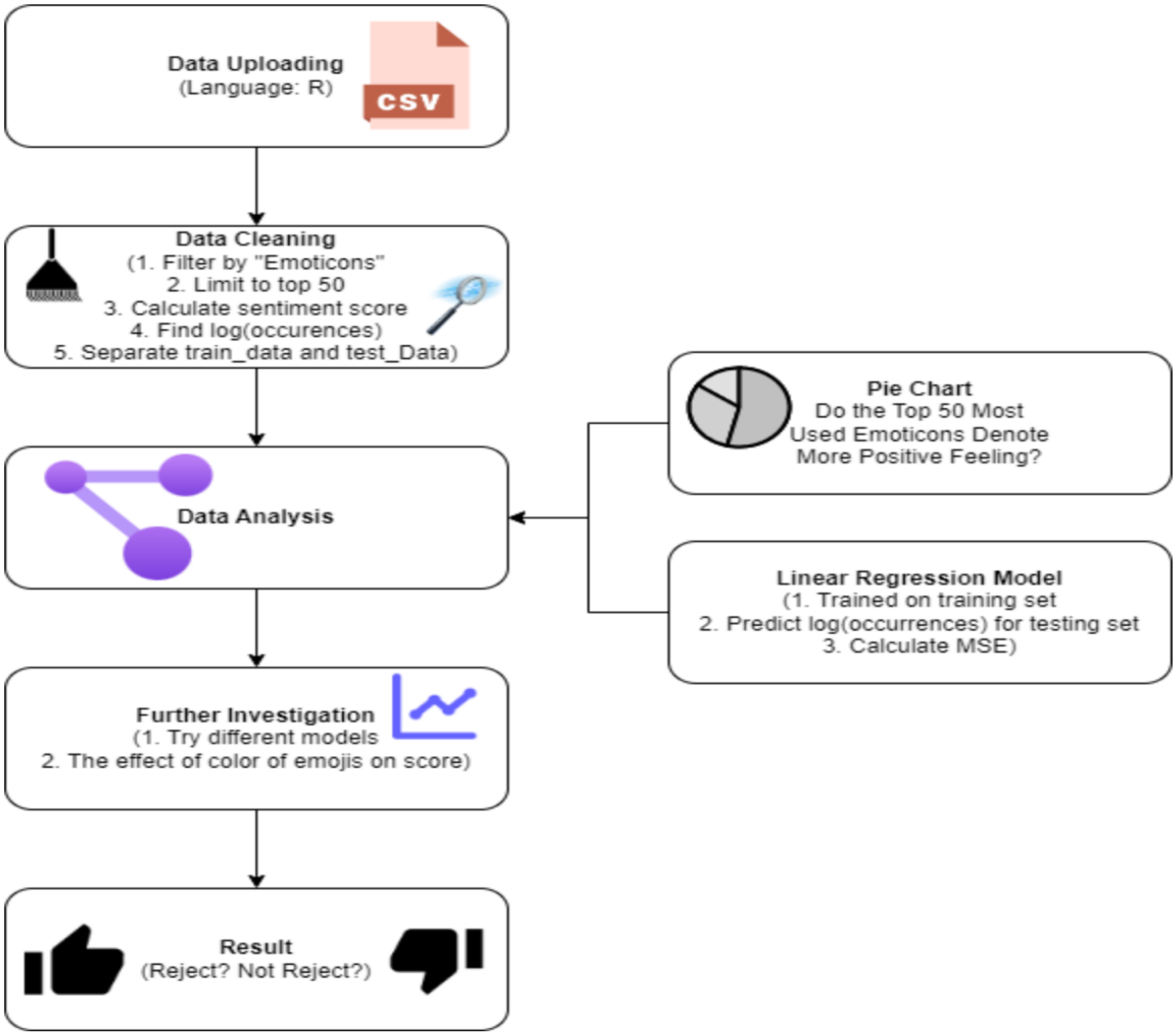
This graph shows the results of the linear regression analysis on the training data. The x-axis has the sentiment score associated with the test value, and the y-axis has the log(number of occurrences) for each sentiment score.

```
Call:
lm(formula = occur ~ score, data = train_data)

Coefficients:
(Intercept)        score
      6.400        1.275
```

The output above tells us that the linear regression equation is log(occurrences) = 1.275 * sentiment_score + 6.4.

Analysis Plan:



The first step of our plan is to do the setup necessary for running our code and analysis. For our analysis, we are going to use R and RStudio. We decided to use R for this analysis because our group has more experience with the software and feels more comfortable using R rather than Python.
The next step is to load the dataset into RStudio. We will use the read_csv function to efficiently read in our dataset. Then, we filter the dataset. We ensure that there are no null values present in the dataset that we will be using. This is especially important because we are going to create a model utilizing this dataset.
The third step in our plan is to do the analysis. First, we will create a pie chart to visualize which feelings are portrayed by the top 50 most used emojis. Next, we will do further analysis by creating a model, which is further explained in the paragraphs below.

Model Development Process:

Rather than doing a decision tree model as we had planned in the hypothesis document, we decided to use a linear regression model instead. First, we split the data into two sets: a training data set that contained 80 percent of the data and a testing data set that contained the remaining 20 percent. Then, a linear regression model was created and trained using the training set that was created. After the model was trained sufficiently, we predicted the log(occurrences) of testing data based on their sentiment score.

We evaluated our linear regression model by utilizing the Mean Squared Error (MSE) value. Essentially, if the MSE value is less than 1, the model is accurate, and the closer the MSE value is to 0, the more perfect a model is [2]. The MSE was 0.766, which means that the linear regression model was accurate at predicting values, as the MSE was less than 1.

Our second-to-last step of our analysis is to do further research, if time permits. We would like to utilize other models, such as the decision tree model, to see if we get similar results as the ones returned by the linear regression model. We would also like to see the relationship between the color of the emojis used and the sentiment associated with that emoji.
Finally, we have our results step. This is where we evaluate our results and either reject or accept our hypothesis. We will determine whether the correlation between the popularity of emojis (emoticons) and their sentiment scores is greater than 0.5.
As for our specific quantifiable goal, we will utilize the MSE to measure the accuracy of our model. The closer the MSE value is to 0, the more accurate the model is. We are hoping to get a MSE value of less than 1.

References:

[1] Berengueres, J. (2017, October 1). *Emoji sentiment*. Kaggle. Retrieved November 9, 2022, from https://www.kaggle.com/datasets/harriken/emoji-sentiment?select=ijstable.csv

[2] W. Rowe, “Mean Square Error & R2 Score Clearly Explained,” *BMC Blogs*. https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/#:~:text=There%20is%20no%20correct%20value