# Emoji Popularity Analysis Using SEMMA Methodology

Kelly Nguyen*

[1]Charles Davidson School of Engineering,
San Jose State University, San Jose, California

**Abstract**

*The paper aims to conduct a rigorous analysis of the factors influencing the popularity of emojis. Utilizing the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, a Linear Regression model was employed to predict the popularity rank of emojis. The model explained approximately 35.7% of the variance in the popularity ranks, providing moderate predictive power.*

## Introduction

Emojis have increasingly become a ubiquitous aspect of contemporary digital communication. Originating as simple pictographs, they have evolved into a complex language of emotional and contextual symbols that supplement and, at times, even replace conventional text. Their proliferation has transformed them from mere digital curiosities into a substantial topic worthy of scholarly inquiry. This transformation raises questions about the factors that drive the popularity and usage patterns of emojis in modern digital lexicon. Understanding these factors can provide valuable insights into broader linguistic trends, social norms, and even the psychological inclinations that influence human communication in a digital age.

The core aim of this study is to identify and analyze these influencing factors through a systematic, data-driven approach. To achieve this, we follow the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology for data science, a robust and systematic framework designed to guide analytical projects from conception to conclusion.

- **Sample**: The initial phase involves gathering a representative sample of data. In the context of this study, we curate a dataset of digital conversations containing emojis, ensuring a balanced representation of various platforms, demographics, and themes.
- **Explore**: This phase entails the initial exploration of data through descriptive statistics and visualization techniques. It serves as a preliminary analysis to detect any apparent trends, gaps, or anomalies that might warrant further investigation.
- **Modify**: Based on the insights gleaned from the exploration phase, the data may be transformed or modified to facilitate modeling. This could include tasks such as data cleaning, normalization, and feature engineering.
- **Model**: This is the crux of the analysis, where various statistical and machine learning models are applied to the modified data to extract actionable insights. For this study, we employ a range of models to determine the factors that significantly influence emoji popularity.
- **Assess**: The final stage involves evaluating the effectiveness and reliability of the models by employing various validation techniques. The insights are then synthesized into a cohesive conclusion, and recommendations are made based on the study's findings.

By adhering to the SEMMA methodology, this study ensures a disciplined and replicable approach to exploring the enigmatic yet increasingly relevant world of emojis. Our research aims not only to contribute to the academic understanding of digital communication but also to offer practical implications for industries like marketing, mental health, and social computing.

**Data Source** The dataset used in this study was sourced from Kaggle and contains 1,549 entries of emojis along with their popularity ranks, categories, and other features.

## Data Structure

In order to conduct a comprehensive analysis, we utilize a meticulously curated dataset that features seven distinct columns, each providing unique facets of information about emojis. These columns are as follows:

- **Emoji**: This column contains the actual emoji character. These are the visual representations

*Corresponding author: kelly.nguyen01@sjsu.edu
**Received:** September 21, 2023, **Published:** September 21, 2023

that are subject to this study, ranging from faces and animals to objects and symbols.
- **Rank**: A numerical index is assigned to each emoji to denote its rank in terms of popularity or frequency of use. The rank serves as a quantifiable measure to assess the relative prominence of each emoji in the dataset.
- **Year**: This column captures the year in which the emoji was introduced. The inclusion of this temporal information allows us to study trends over time and to understand how the popularity of emojis might be influenced by their age.
- **Category**: Emojis in this dataset are grouped into various categories for easier classification and analysis. These categories could range from "emotions" and "animals" to "food & drink" and "travel."
- **Subcategory**: Within each category, emojis are further classified into subcategories for more granular analysis. For example, the "emotions" category might have subcategories like "happy," "sad," and "angry."
- **Hex**: This column provides the Unicode hexadecimal representation of each emoji. This standardized coding system for characters offers a machine-readable format that can be useful for automated analysis and data manipulation.
- **Name**: The last column contains the official name or description of each emoji, providing a textual representation that supplements the visual character.

During the data preparation phase, an extensive quality check was performed to ensure the integrity and completeness of the dataset. Remarkably, no missing values were identified, confirming the dataset's reliability and robustness for the analysis.

This well-structured dataset, free from missing values and endowed with multiple features, provides a solid foundation for conducting the study. Adhering to SEMMA methodology, the dataset is first sampled to ensure representativeness, followed by exploration, modification, modeling, and assessment.x

# Methodology

## Exploration

**Visualizations** Our initial visualizations revealed notable trends in emoji introduction and category frequency. We found that a significant majority of emojis were launched between the years 2010 and 2015. This period seems to coincide with the rapid expansion of smartphone usage and the proliferation of social media platforms, making it a critical era for the integration of emojis into digital communication. Interestingly, among the various categories, the "Flags" category surfaced as the most frequent, potentially signifying its broad applicability in various conversational contexts.

## Modification

**Data Cleaning** During the data preparation phase, we encountered minor inconsistencies that could have significantly impacted the modeling stage. These included issues like leading spaces in column names, which were meticulously identified and rectified. Such data hygiene practices are critical for ensuring that subsequent analytical steps proceed without error or bias.

**Feature Engineering** To enhance our analysis, new features were engineered based on existing variables. One such feature is the length of the emoji name, which could potentially correlate with its ease of understanding or popularity. Additionally, an indicator variable was created to flag whether the emoji's name contains the word "face," as face emojis often serve specific emotional expressions and might have a different level of popularity compared to other types of emojis.

## Modeling

**Model Selection** After considering various machine learning algorithms, we opted for a Linear Regression model due to its straightforward interpretability and mathematical simplicity. This choice allows us to easily dissect the model later for better understanding of the factors affecting emoji popularity.

**Training and Validation** The selected Linear Regression model was trained on a carefully chosen subset of the data, while validation was performed on another distinct subset to prevent overfitting. The features used for training and validation include the 'Year' of introduction, the engineered 'Name length', and the indicator variable 'Contains face'.

## Assessment

**Model Metrics** The model's performance was evaluated using multiple metrics, namely Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared ($R^2$) Score. The computed MSE was 129779.17, the MAE stood at 305.38, and the $R^2$ Score was 0.357.

**Interpretation** The model successfully explained approximately 35.7% of the variance in the ranking of emojis. While this suggests a moderate level of predictive power, it also implies that there are other

unconsidered variables or complexities that might be influencing the popularity ranks of emojis. Therefore, although the model provides a reasonable basis for understanding trends, further refinements and feature inclusion may be necessary for a more comprehensive understanding.

## Summary and Findings

The core objective of this study was to explore the factors influencing the popularity rankings of emojis, aiming to provide insights into linguistic trends and digital communication patterns. Utilizing a Linear Regression model, the study was able to explain approximately 35.7% of the variance in emoji ranks. While this is a notable achievement, it does illuminate areas for further refinement.

The moderate level of predictive power achieved is constrained by several factors. Firstly, the simplicity of the Linear Regression model itself could be a limiting factor. While simple models are often easier to interpret, they may lack the complexity needed to capture nuanced relationships in the data. Secondly, the dataset used for the study might lack additional features or contextual information that could shed more light on emoji ranks, such as geographical or platform-specific usage data.

## Future Work

Looking ahead, there is considerable scope for further research and development in this domain. One immediate avenue is the exploration of more advanced machine learning models that can capture complex, nonlinear relationships in the data. Models such as Random Forests, Gradient Boosting, or Neural Networks may offer higher predictive accuracy and could be evaluated in future studies.

Moreover, the process of feature engineering has the potential to significantly elevate the model's performance. Variables like the emotional valence of an emoji, its frequency of use in different types of communication (e.g., social media vs. formal emails), or even cultural factors could be synthesized and included in future iterations of the model.

Finally, while the present study successfully applied the SEMMA methodology in breaking down the various phases of the data science pipeline, future work could also focus on incorporating other frameworks or techniques for a more comprehensive analysis. This could involve comparing SEMMA with other methodologies like CRISP-DM for a holistic understanding of the problem space.

## References

[1] Kaggle, *Emoji Dataset*, n.d., Retrieved from `https://www.kaggle.com/`.

[2] Wikipedia, *Emoji*, n.d., Retrieved from `https://en.wikipedia.org/wiki/Emoji`.

[3] SEMMA, *SEMMA Methodology for Data Science*, n.d., Retrieved from `https://www.sas.com/en_us/insights/analytics/what-is-data-science.html`.