# Project Proposal

Robert Ivill 46012819

Can we identify patterns between the style of beer and its rating across multiple dimensions of its brew?

This project aims to explore patterns between various beer styles and their ratings across multiple dimensions of its brew, using the RateBeer dataset. This research question aims to pique the interest of both beer enjoyers, stakeholders within the beer industry, and those with a love for big data. By leveraging matrix factorization algorithms, I intend to find hidden links between how a person perceives a beer and its style. This will provide valuable information on which characteristics contribute most to a consumer's satisfaction, in turn guiding product development and enhancing the beer experience worldwide. Dask's capabilities to handle large data with its parallelism will allow for efficiency throughout the process. The design will include loading, preprocessing, factorization, analysis, and visualization of the data in relation to the research question.

Pursuing the answer to this research questions is important to me as a data scientist and beer enthusiast. I am eager to see how people tend to rate different style beers across their different features. I hope to uncover where some styles of beer fall short or shine and see what the beer community has preferences to and biases against. This will allow me to explore the intricate nuances that come with beers and find valuable insights into what it takes to make a great beer. This is important not only to the beer community, but the shareholders within the industry, as insights like this can guide product development and enhancements of the beer drinking experience globally.

It is important to provide a background to the features that are apart of this dataset and what the beer-related jargon means. Beer enthusiasts often use multiple aspects of a beer to assess their rating. Those included in the dataset are taste, appearance, aroma, and palate, however, these are not the only factors that contribute to a final rating. Beer style relates to the overall characteristics of a beer, how it's made, and where it's from. Regarding the algorithms, we will use a technique called matrix factorization that will break down each style of beer into its own matrix of ratings across the features. We will have to make use of mapping, reducing, and grouping the data by these features to produce factorized matrices. Using distance measurements will allow us to see the vast differences and similarities between different styles of beer.

The chosen research question is: "Can we identify patterns between the style of beer and its rating across multiple dimensions in the RateBeer dataset?" This research question is relevant for the RateBeer dataset as it contains comprehensive reviews with aspect specific ratings for various beer styles. By exploring these patterns between beer styles, we can gain insight into which characteristics contribute to satisfaction for consumers. Implementing matrix factorization algorithms will help answer this question by decomposing the rating data to uncover factors that represent hidden patterns or features within the dataset. This approach will allow us to find common traits among the beer styles and find hierarchies across beer styles.

For this project, I will be using Dask to play the role of handling the large-scale dataset effectively. Dask will be used for the initial loading and preprocessing of the data, with its parallel computing capabilities being used to handle the large size. The data flow will include constructing a user-item matrix in Dask Dataframes, followed by applying matrix factorization techniques, such as Singular Value Decomposition, to decompose these matrices. These decompositions will be analyzed to identify the patterns between beer styles and their ratings across multiple dimensions.

The program design will include components for the loading, preprocessing, matrix construction, factorization, analysis, and visualization. This will leverage Dask's parallel processing powers. The sequence of events in the design will be loading the dataset with Dask, preprocessing the data to handle missing values and outliers, constructing user-item matrices into Dask dataframes, applying analysation techniques to compare by beer style and finally using NumPy and Matplotlib to visualize the results. Potential difficulties with this sequence of events may come from managing the distribution of resources efficiently, task scheduling when applying parallelism, and ensuring that libraries such as Dask and Matplotlib are compatible throughout the design process. Despite this, Dask will enhance the scalability and performance of the analysis, meaning that patterns within the RateBeer database can be explored quickly and easily.

References:

*Beer styles*. BeerAdvocate. (n.d.). https://www.beeradvocate.com/beer/styles/

Guest, C. (2018, April 2). *What is a beer connoisseur?*. The Beer Connoisseur®. https://beerconnoisseur.com/articles/what-is-a-beer-connoisseur

Team, G. L. (2024, February 23). *Matrix factorization explained: What is matrix factorization?*. Great Learning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/matrix-factorization-explained/