1. Background/significance of the research and research question (**4** points). You should describe some background information about your data and why it is important to perform classification or regression. Then, state your research question(s), which must be clearly related to your data and the methods to be used.

Research question: The goal of this project is to predict the future popularity of beers given certain factors. We will be analyzing a dataset to determine which keywords, type, brand, aroma, style, appearance, taste and so forth affect what beers people prefer. Using this will we create predictions on beer popularity.

We are gonna treat our data as continuous, Try to do classification on the names of the beer. We can do k-nearest neighbors, might be useful. We can use methods from chapter 8.

Beer is one of the most popular alcoholic beverages in the world, with a global market worth billions of dollars. Understanding what factors contribute to the popularity of certain beers can be crucial for breweries and other businesses in the industry. With the rise of craft beer culture and the increasing number of beer options available, predicting what beers will be popular in the future can be a valuable tool for businesses looking to stay ahead of trends and improve their product offerings.

Our goal is to use a dataset of beer reviews, to answer the following research question: What factors contribute to the popularity of different types of beers, and can we create a model to predict the future popularity of beers given these factors? Specifically, we will be analyzing the relationship between beer attributes such as keywords, type, brand, aroma, style, appearance, and taste, and the overall popularity of a beer as determined by user reviews. While treating our dataset as continuous we will use classification, K-nearest neighbors and tree based methods to develop models that can predict future popularity based on these attributes.

2. The methods used to obtain and analyze the data (**4** points). You should describe the methods used. You are required to use at least two different regression, clustering, and/or classification models (e.g., logistic regression and LDA). If you are familiar with machine learning programs in other programming languages, you may use those as well, but see notes below for more details.

The dataset used for this project is a collection of user reviews and ratings for a myriad of beers. The dataset includes various attributes of the beers such as style, aroma, taste, appearance, and overall rating. We will be using this data to predict the popularity of beers based on these attributes. The original dataset contains over 1.5 million reviews in total. However, working with such a large quantity of data is computationally expensive, and as we preprocessed our data, we were able to reduce the number of reviews to around 56,000.

To preprocess the data, we first removed any unnecessary columns to our research such as time of review, then converted categorical variables into dummy variables to be used in our models. Additionally, we wanted to reduce the size of our dataset without compromising the accuracy of a beer’s reviews by removing a subset of a given beer’s reviews. Therefore, we elected to take an average of all reviews for a unique beer in each category and use these values to create one data point. Thus, for each of the 56,857 unique beers in the dataset, there is one review that is representative of all given reviews in the original data. This reduces the size of our data set from 1.5 million to 56,857, which is somewhat more manageable for conducting analysis.

We further separate each beer by style. There are 104 unique styles present in the data set, and we are creating graphs and performing analysis within each style of beer. One benefit to this is that our graphs will be easier to interpret (and also feasible to graph in R for some computers), and more significantly, this helps ensure our analysis is useful in application. For example, if a dark beer such as a Russian imperial stout has a high overall review rating, it would make sense to compare its taste and aroma to other dark beers of the same style to draw inferences about why it is well-liked instead of comparing it to a light lager beer, which often has a completely different flavor profile and should not be held to the same standards. Additionally, some beer drinkers may have strong preferences for beer styles, and to ensure they are recommended a beer they like, it is best to recommend beers within a style they enjoy.

For classification, we will be using two different methods: K-nearest neighbors (KNN) and tree-based methods. KNN works by identifying the k closest data points to a new observation and using the most common class among those k neighbors to predict the class of the new observation. Tree-based methods, such as decision trees and random forests, work by recursively splitting the data based on the most informative attributes until the data is partitioned into homogeneous subsets.

We will also be using clustering methods to identify any natural groupings in the data. Specifically, we will use K-means clustering to group beers based on their attributes. K-means clustering works by assigning each observation to the nearest cluster center and iteratively updating the center until convergence.

Overall, we will use a combination of classification and clustering methods to develop models that can accurately predict the popularity of beers based their given (part attributes.