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.

Using a dataset of beer reviews from BeerAdvocates, our goal was to create our own beer recommendation model. In the final product, the user would be able to input the name of a beer that is within the dataset and receive a list of recommendations of similar beers based on our model. This model takes factors such as brand, aroma, tyle, appearance, and taste into account to make a list of recommendations; by default, the model gives five recommendations, though more or less may be specified. While treating our dataset as continuous, we will use classification and K-nearest neighbors to develop our model.

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 45,000.

To preprocess the data, we first removed any unnecessary columns to our research such as time of review and the reviewer’s user ID. 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 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 experimented with further separating each beer by style. There are 104 unique styles present in the data set, and we attempted to perform analysis within each style of beer. One benefit to this is that our graphs would be easier to interpret (and also feasible to graph in R for some computers), and more significantly, this would help 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 would be best to recommend beers within a style they enjoy.

However, we elected to not use this method for several reasons. Primarily, we wanted a simpler model, and the method above would have created a lot more complexity than the scope of the project allows. Additionally, we had some concerns about the usefulness of only recommending beers within the same style; while this would be simple, we did not feel as if it was truly creating a recommendation based on the characteristics of the specific inputted beer, which was our goal.

Instead, we ended up using K-means clustering with k=5 on the whole dataset, then assigned the cluster labels to our data. Using a for loop, we tested values of k from 1 to 100, and then selected the value of k at which the within-group sum of squares decreased by less than 5% as k increased by 1; this is essentially the same as selecting the “elbow” point from a graph. After deciding on k=5, we performed K-means clustering on the data. We then wrote a function that calculated the Euclidean distance between the input beer and beers within the same cluster. The function returned the names of the beers that have the shortest Euclidean distance to the given input beer.

In addition to K-means clustering, we used a Lasso regression to identify which predictor variables influenced the overall review score the most.

DISCUSS THE LASSO METHOD, AND THEN WE’RE DONE