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.

First, we will preprocess the data by removing any unnecessary columns and handling missing values. We will then convert categorical variables into dummy variables to be used in our models.

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.