Naina Wodon

Data 1030: Project Proposal

Introduction

Whiskey is a popular beverage consumed in the United States, with a number of different styles and types. Due to the large diversification of the whiskey market, the beverage can be sold at many price points.

We hope to determine which types of whiskey will be both profitable and marketable. Our goal is to identify the types of whiskeys customers should drink and distilleries should brew using a dataset called *2,2k_ Scotch Whiskey Reviews* which we obtained from Kaggle. We believe that by using this dataset, we will be able to determine whiskey trends and preferences and thus adequately serve the whiskey market.

Our method includes creating two predictive models that can help new whiskey drinkers decide what whiskeys they should try depending on their price point and taste preferences. These models will be particularly useful to young adults who are interested in trying whiskey or adults who are interested in increasing their whiskey collections based on the price of a bottle.

Natural language processing will also be used on the description portion of a whiskey label to determine the most popular whiskey flavors. Whiskey distilleries may be interested in the results of NLP analysis as it could inform their decisions on which whiskey flavors to emphasize. In a similar vein, we will also determine what the key words on a whiskey label are and which descriptions are correlated with higher whiskey review scores and prices. Through an analysis of the data set we hope to better understand current whiskey trends and what drives positive reviews.

As part of the initial analysis, we will determine which flavors occur most frequently in highly rated whiskeys and what the average price and review score for each category of whiskey is. We will subsequently build a whiskey model that predicts whiskey price, and a model that predicts whiskey review score. When building these two models we will use cross validation techniques and also try to see if we can incorporate some of the word to vec analyses in our model.

Data Understanding

The data, which consists of well-known whiskey bottles from large distilleries was from *2,2k_ Scotch Whiskey Reviews*, last collected in June, 2018; it included 2,248 different whiskeys with 6 different whiskey styles. The dataset contained 7 different features “ID”, “name”, “category”, “review point”, “price”, “currency” and “description”.

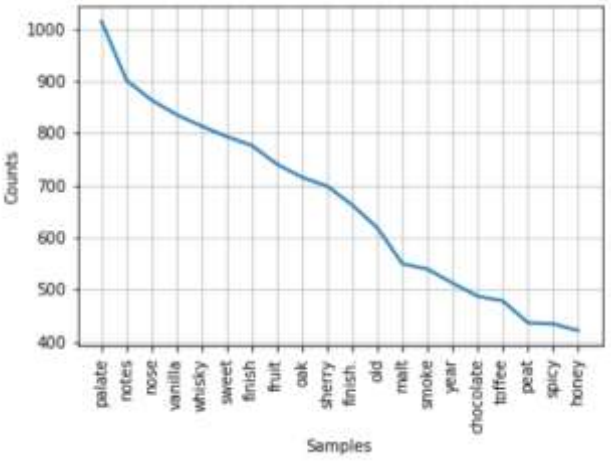
Data Preparation

Note that we removed the variable ID from our dataset because it is unrelated to a whiskey's physical properties and therefore would not affect our analysis. Currency was also dropped from the dataset because all the whiskey prices were in dollars. The price variable was encoded using the standard scalar transformation because there are some very expensive whiskeys in the dataset. Review point was encoded using the MinMaxScalar because the feature is bounded between 0 and 100.

Whiskey type which is defined under the variable "category" was hot encoded. The features whiskey name and description were not hot encoded as each observation had a unique whiskey name. These variables will be analyzed using natural language processing. Once these variables have been processed, the occurrence of certain words will be hot encoded.

From a business perspective, the name and description of a whiskey is an influential part of the branding. Therefore, by analyzing these two features we hope to glean some insightful information on what are the most common whiskey trends at the moment.

To analyze the frequently used words in the names and descriptions of each whiskey, we will break each name and description down to a list of its words and compiled all of those into one document. We will then remove the punctuation and the words that are considered common in the English language (e.g. "the," "a," "and"). The remaining words we are left with will contain exclusively title and description words for each whiskey. From this document we will generate the term frequency information for each word, and plot the words with size by frequency in a word cloud. We will also create word vectors and utilize these vectors in our predictive models for whiskey price and review score. Through an initial analysis, we were able to determine the most frequently used words in the description section of a whiskey label (outputted to the left).



The dataset used for this project has been used in several other projects. One project used k-means analysis to determine what are the main characteristics for each whiskey type, and tried to determine if there was a correlation between whiskey price and review score. Another project used natural language processing to build different models to classify the whiskeys by type. The data was also used in a sentiment analysis project which categorized reviews by sentiment.