Scotch Whiskey Flavor Profile Classification

By William Roberts - HarvardX Data Science 2019

1/28/2019

Also hosted on a dedicated github repository (https://github.com/RobertsData/Data-Science-Projects.git)

Introduction

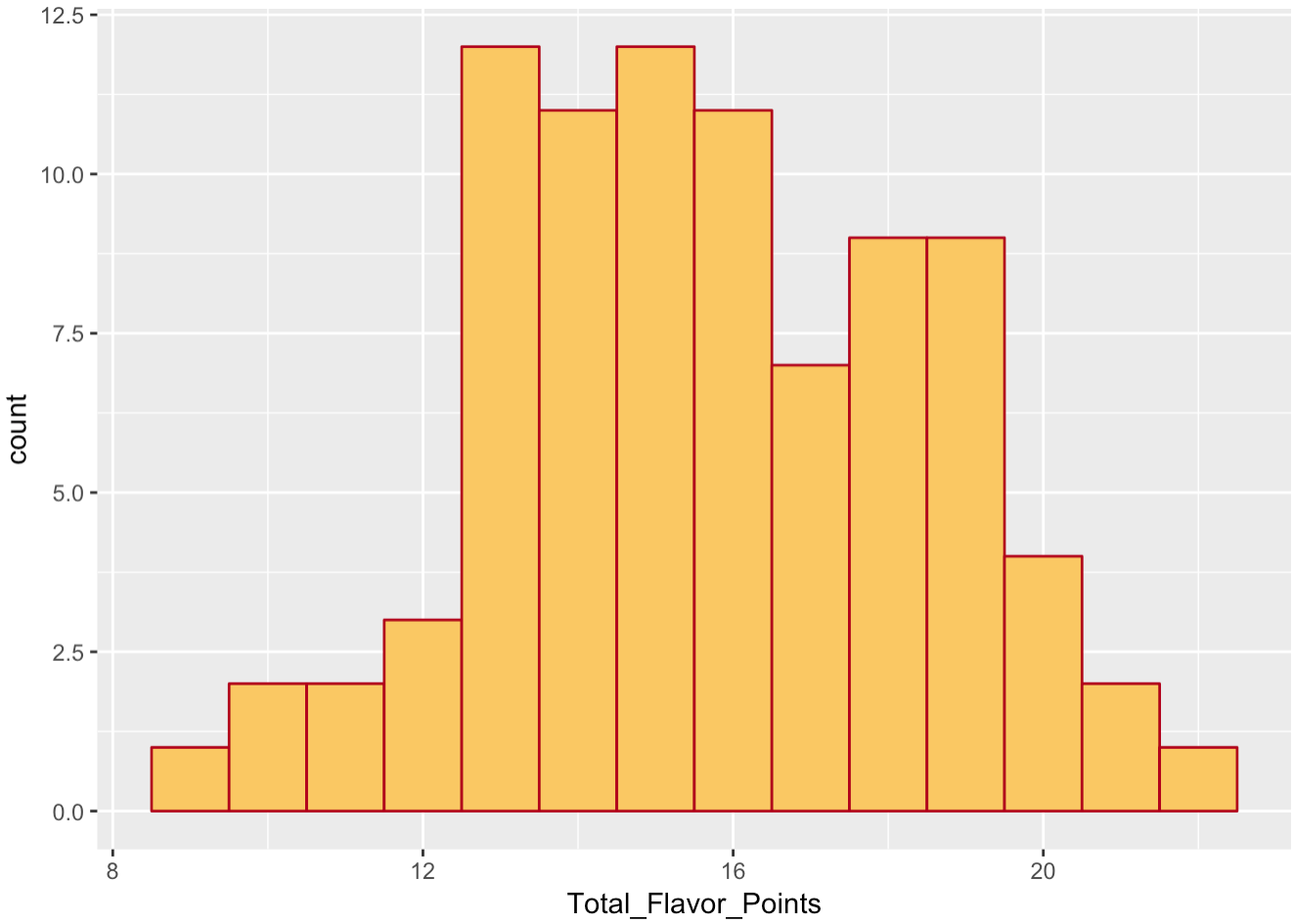
Scotland is famous for its tradition of distilling high-quality whiskey, often called “scotch” outside of the UK, encompassing several unique methods and distinctive flavors. Many distilleries producing scotch have been operating since the 19th or even 18th centuries and have established lucrative, internationally recognized brands.

This analysis uses a dataset of flavor profiles from 86 single malt (non-blended) whiskies produced by different Scottish distilleries. Within the dataset there are 12 flavor attributes, such as ‘Honey’, ‘Malty’, ‘Nutty’ etc listed for each whiskey on a scale from 0-4 according to the strength of each flavor. The goal of this analysis is to gain quantitative insight into which flavors contribute to giving particular types of whiskey their distinct flavor profiles.

The dataset was compiled originally from the book ‘Whiskey Classified: Choosing Single Malts By Flavor’ by David Wishart. [1] Each whiskey in the dataset is a single whiskey chosen as the most representative for each distillery, though distilleries typically produce several different whiskies of different styles.

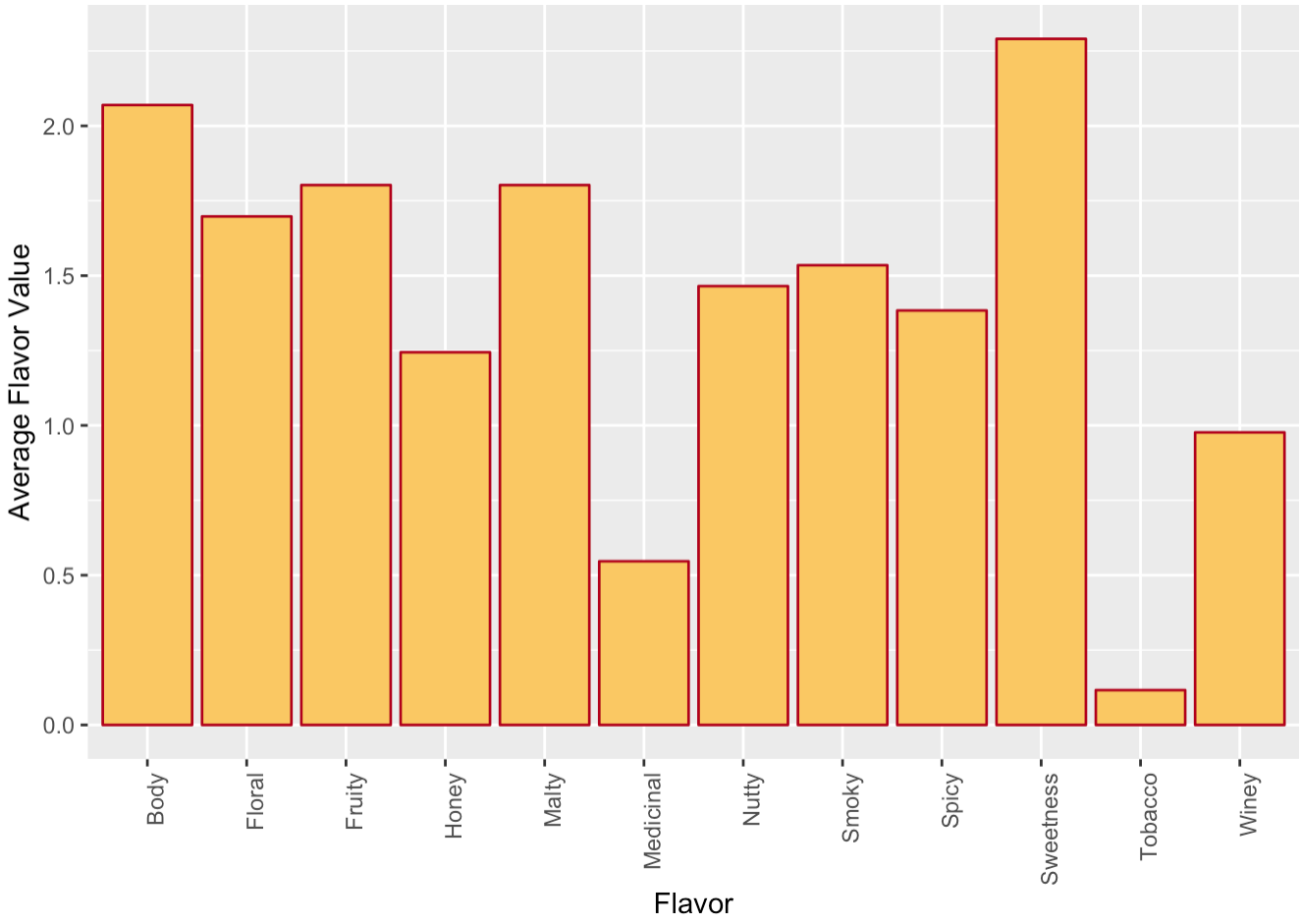
Methods

To begin we should examine how flavor is distributed throughout the dataset. A good way to do this is to examine the distribution of total flavor points for each whiskey.



We can see how the distribution of total flavor points is approximately normal, or perhaps slightly bimodal, with relatively few distilleries producing whiskies that have very little or very strong flavor (high point total).

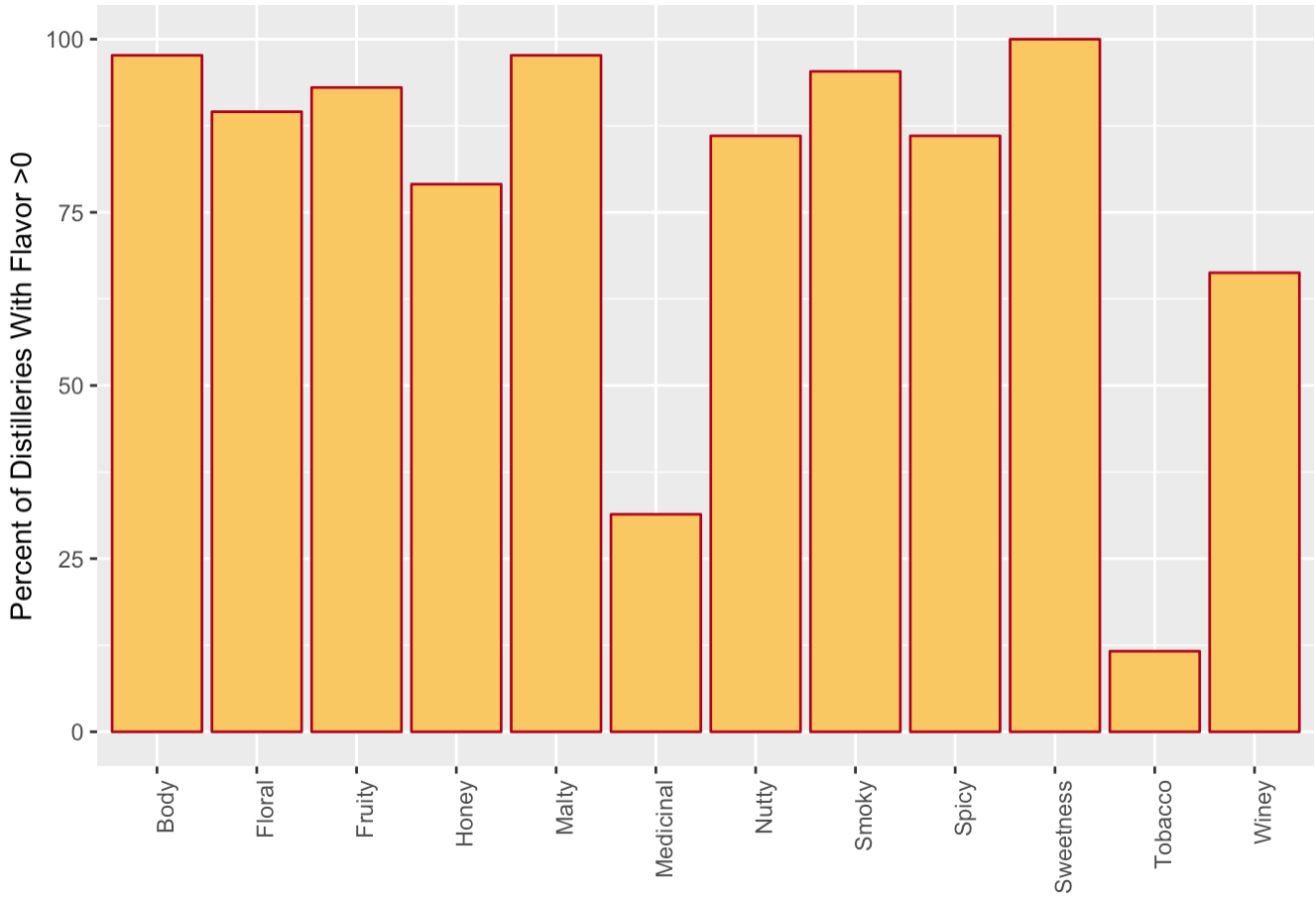
Next we can examine the average values for the 12 flavors used the dataset. This will tell us which flavors tend to be higher or lower than the others across the dataset as a whole.



Note how 'Sweetness' and 'Body' have the highest average rating while the 'Medicinal' and 'Tobacco' flavors tend to be much lower.

'Body' is a difficult flavor attribute to explain but is genrally used to describe how intense or complex the flavor is. The 'body' of a whiskey can be described as how it "fills the mouth" when first tasted and generally becomes stronger with increased barrel aging.[2]

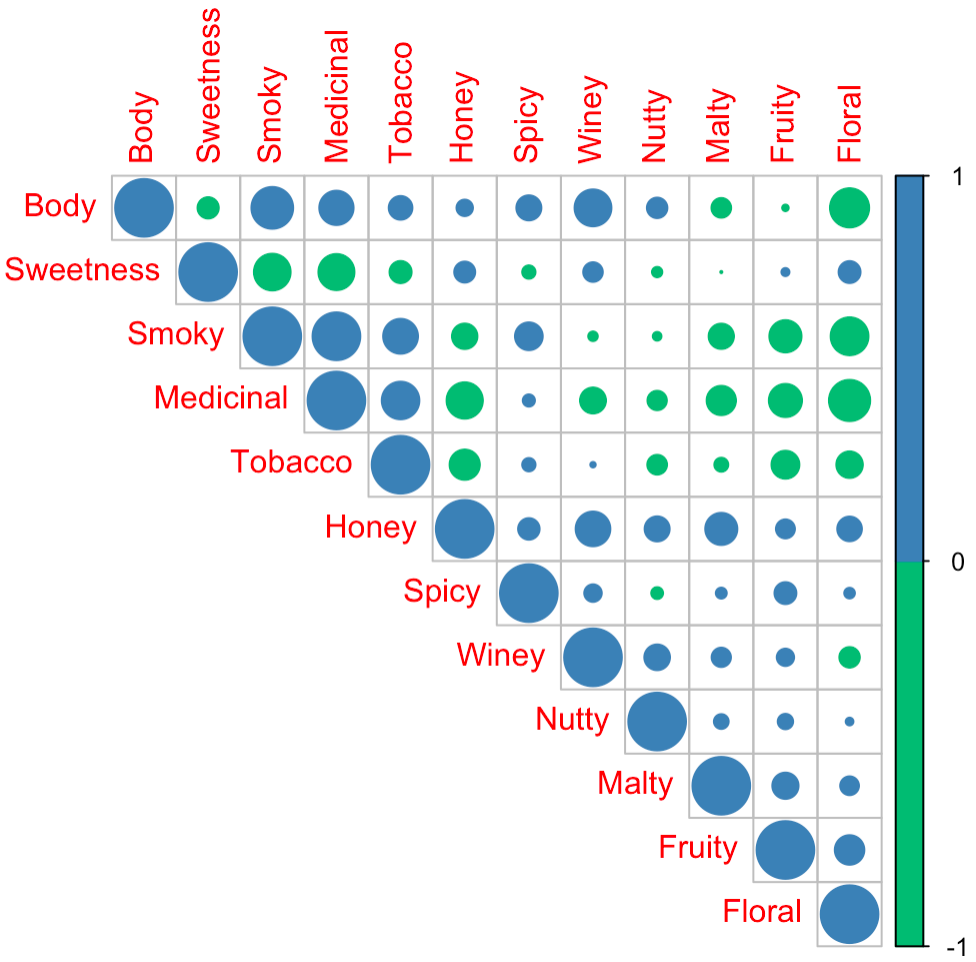
Many whiskies have at least one of the 12 flavors scored as 0, where the particular flavor is not at all percetible. By looking at which flavors are most prevalent within the dataset (most non-zero values) we can see how common they are and how much they contribute to the flavor profile of whiskies in the dataset.



For most of the 12 flavors typically 85+% of whiskies have a value of at least 1. The flavor attributes ‘Sweetness’, ‘Malty’ and ‘Body’ almost always have a value greater than 0, while more than half of whiskies score 0 for the ‘Medicinal’ and ‘Tobacco’ flavors. As we will see later in the analsysis, this is because these particular flavors are mostly limited to a small number of whiskies with a distinct flavor profile.

Certain aspects of the flavor profile are highly correlated with each other, either positively or negatively. A helpful way to visualize the relationships between aspects of flavor profile is to generate a correlation matrix showing how each of the 12 flavors is correlated with all of the others. This is easily done through the corrplot R package.

Each row/column combination in the plot below displays the size of the correlation coefficient as different size circles. Strongly correlated flavors will tend to vary in the same direction with each other from whiskey to whiskey, either in the opposite (green) or the same directions (blue).

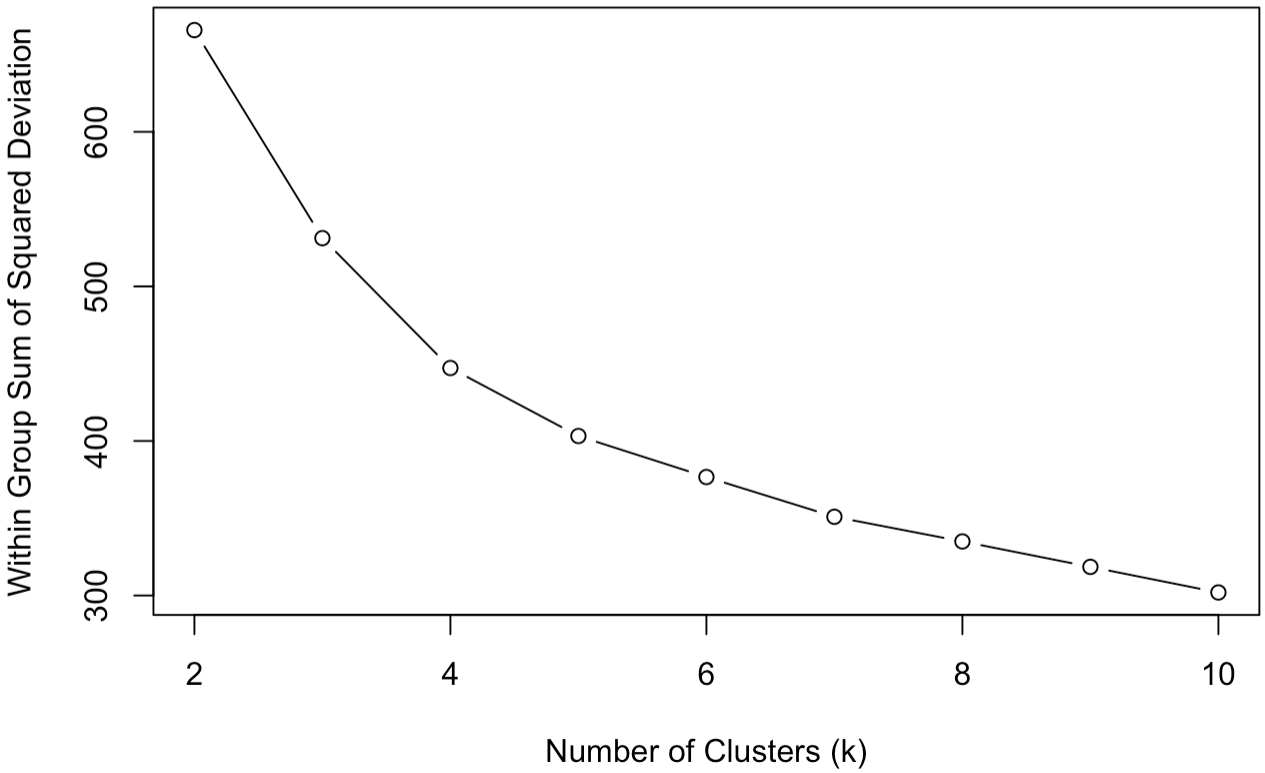


Note how flavors such as 'Smoky', 'Tobacco' and 'Medicinal', while positively correlated with each other, are very negatively correlated with 'Floral' and 'Sweetness' meaning these two sets of flavors are rarely present within the same whiskey.

Now that we have a good understanding of how flavor points are distributed throughout the dataset we can group them into distinctive groups using the k means algorithm (part of the caret package). K means is an unsupervised clustering algorithm that will group the whiskeys into k distinct clusters. Grouping the whiskies into clusters will help us to identify clusters of whiskies with similar profiles that we can easily categorize.

To implement the k means algorithm we need to choose an optimal number of clusters (k). Ideally an optimal value of k would be large enough to separate the whiskies into meaningful clusters while still small enough that diferent clusters will have some significant, meaningful differences. A good metric to quantify this is the total residual sum of squares within each group for different values of k. The sum of squares is calculated as the total squared distance of each flavor from the average value for that cluster.

The plot below illustrates how the variation within groups decreases with increasing k as the data is continually divided into smaller, more homogenous groups. [3]



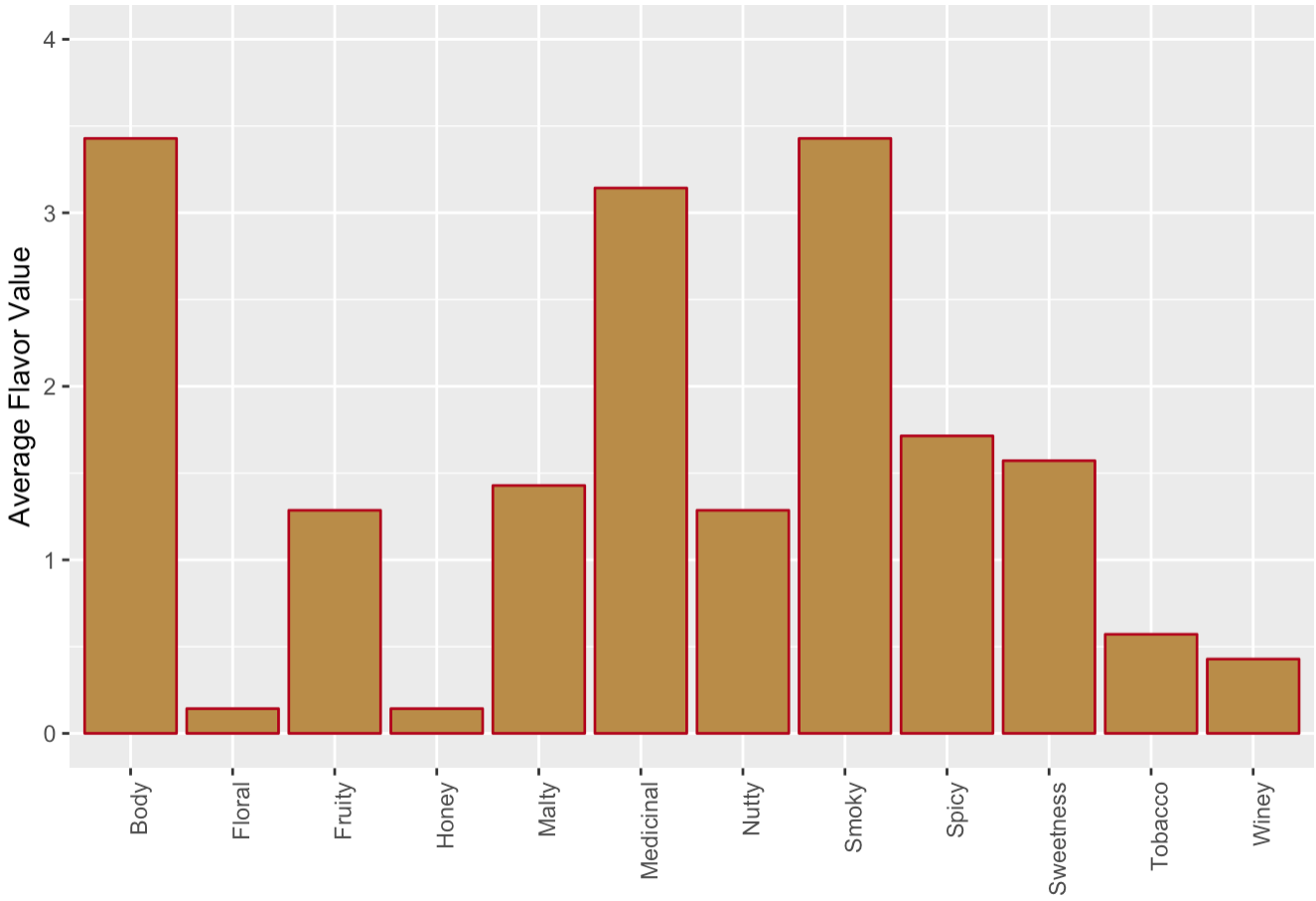
Moving from 2 to 3 to 4 groups greatly reduces the within-group variability. However, values of k greater than 4 do relatively little to reduce the mean squared error. Thus 4 was chosen as the number of clusters in which to divide the whiskeys.

Once the whiskies have been classified and labeled into 4 groups it would be helpful to understand which flavors are most important in distinguishing the groups from each other. One way to accomplish this is by applying a classification algorithm, such as random forest, to the labeled clusters after the groups have been assigned. The variable importance table generated by the random forest classifier will give us some, admittedly imperfect, insight into which flavors are driving the classification into different groups.

Results

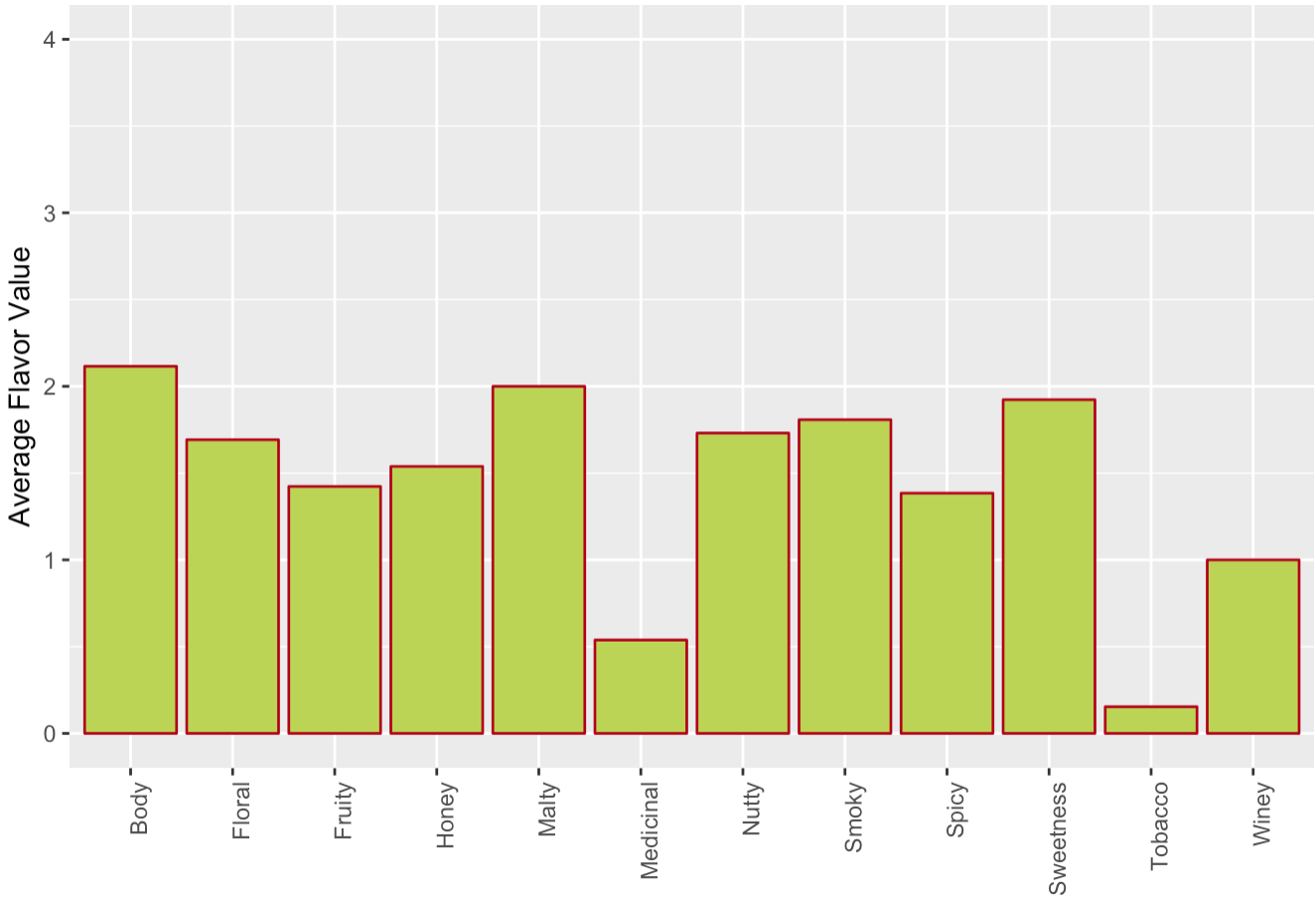
The k means algorithm clustered the 86 distilleries into 4 distinct clusters:

Each of these four clusters has its own distinct flavor profile.

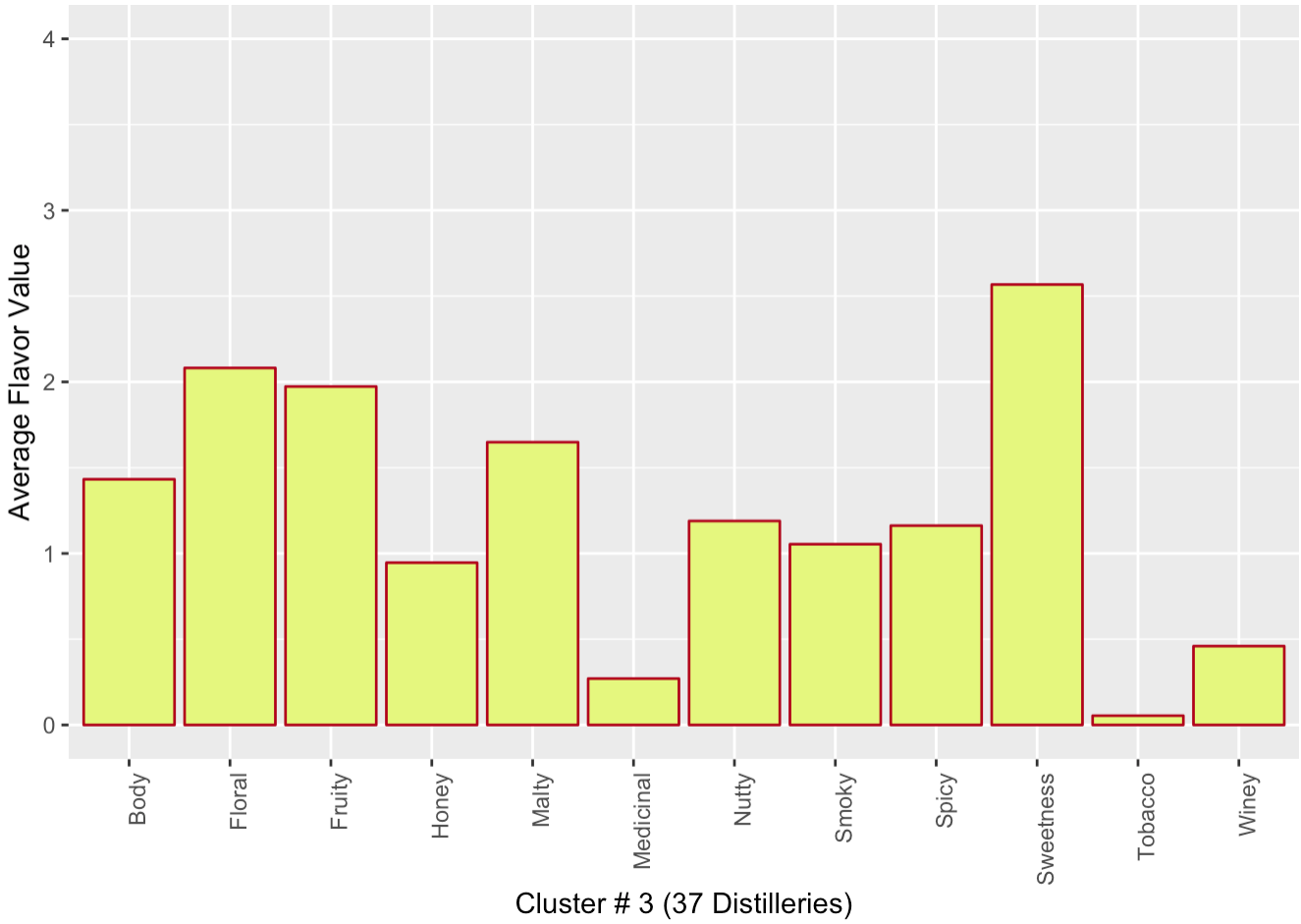


Cluster # 1 (7 Distilleries)

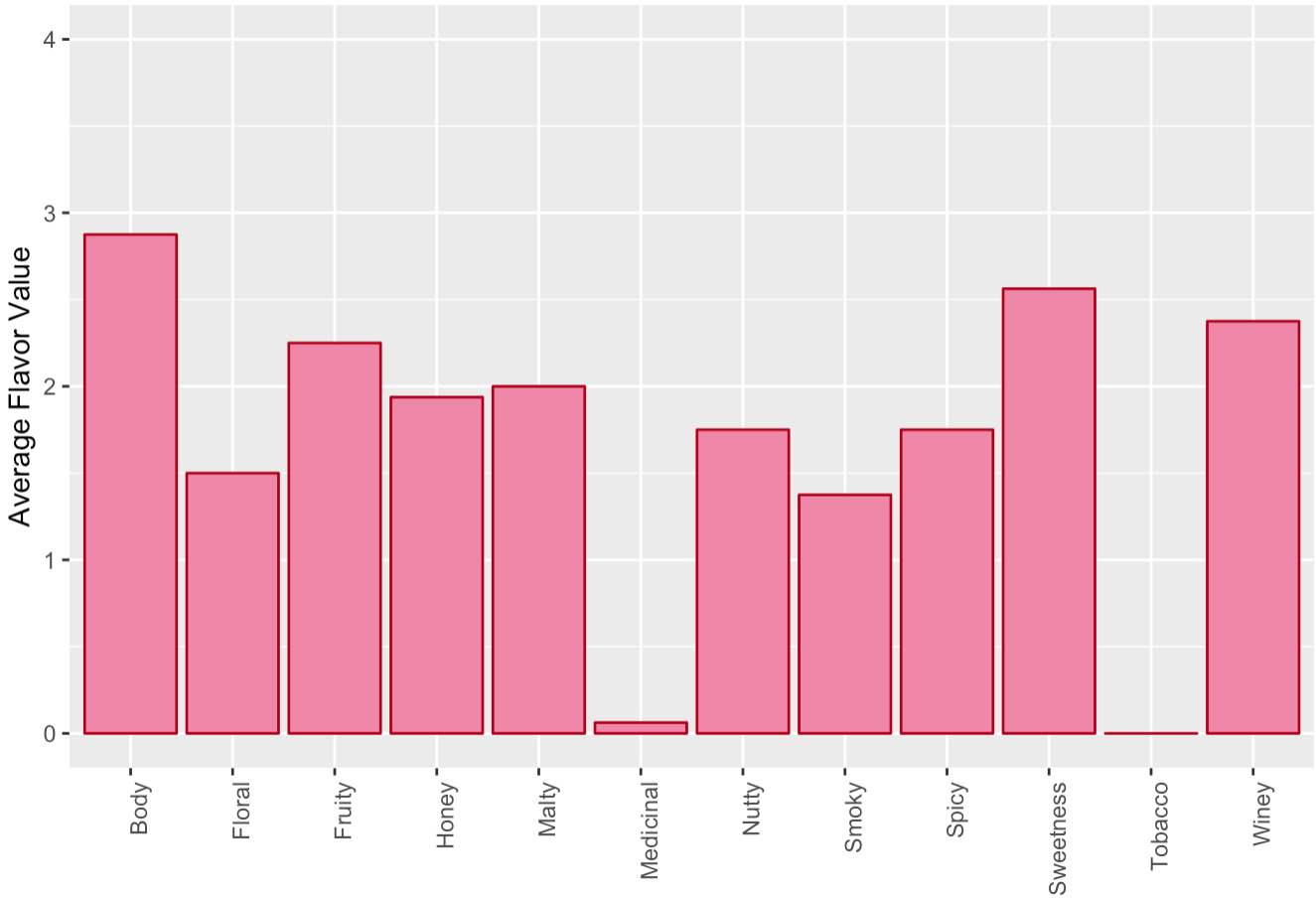
Cluster #1 is the smallest and is made up of smoky, medicinal, high ‘body’ whiskies. Four of the seven distilleries in this group are located on the isle of Islay off of Scotland’s western coast. In this part of the world dried peat was the main source of fuel before modern times and is traditionally used to smoke and dry the malted barley before distilling, giving the resulting whiskey a very strong smoky, medicinal flavor. The most recognizeable brands in this category include Laphroaig and Ardberg, whose distilleries are located within just a few kilometers of each other.



Cluster #2 is made up of more balanced, moderate body whiskies. Some of the more recognizeable brands in this cluster include Highland Park and Dalwhinnie, a distillery nestled in a cold remote corner of the Scottish highlands. Visitors to the Dalwhinnie distillery are offered gourmet chocolates to pair with their whiskies during tasings in order to bring out certain flavors.



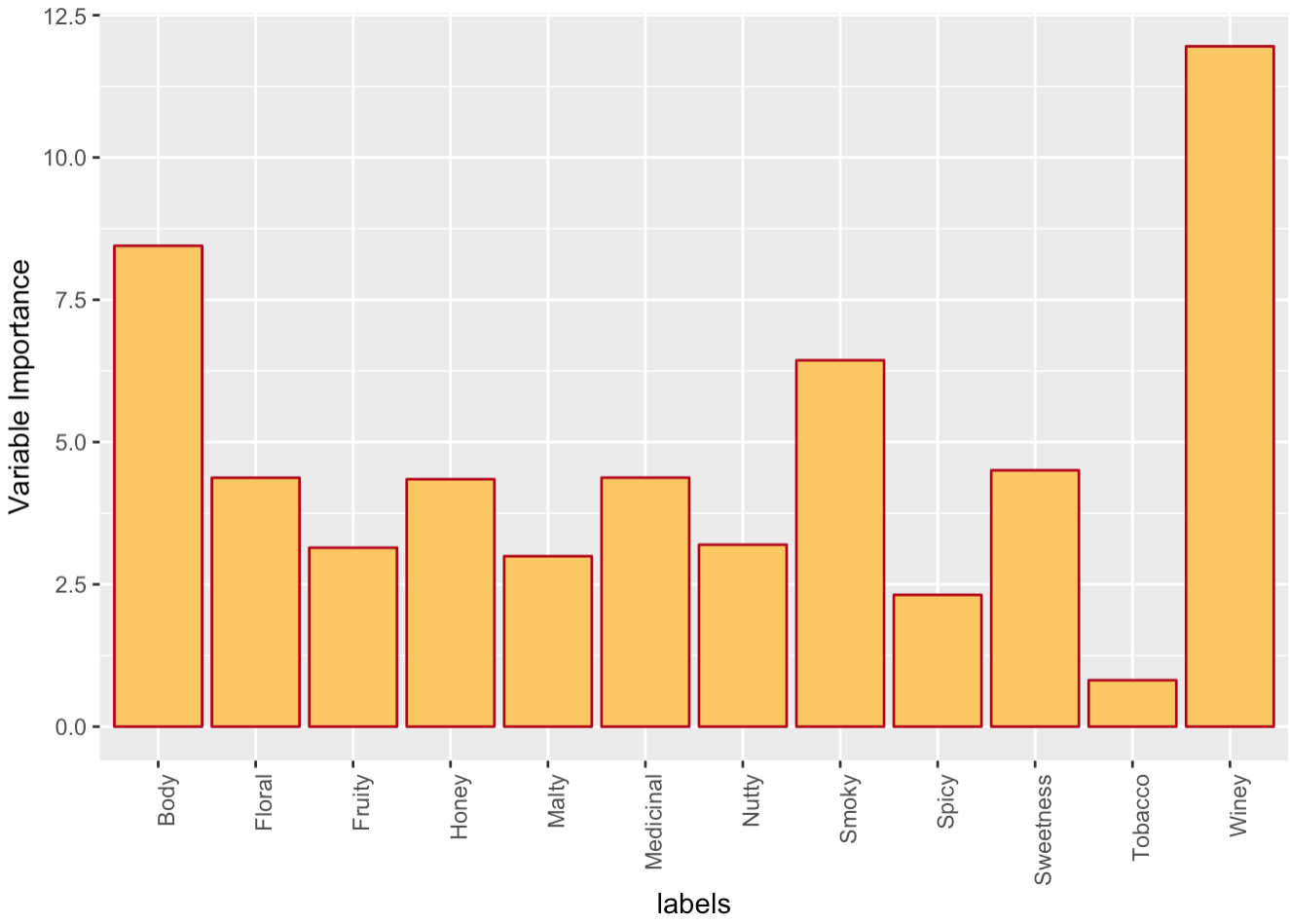
Cluster #3 is made up of sweeter, more floral and fruity whiskies with low body. This is the largest group of whiskies derived from the k means agorithm with 37. Many, though by no means all, of these distilleries are located in the Speyside region of the Northeast scottish highlands.



Cluster # 4 (16 Distilleries)

Finally cluster #4 consists of high-body, very 'Winey' tasting whiskies that still retain a lot of sweetness. The most recognizeable brand from this cluster is Glenlivet, the best selling scotch brand in North America.

Now that we have a good understanding of what makes the clusters distinct from each other we can examine the variable importances from the random forest classifier.



According to our random forest classifier the most important distinguishing flavors attributes are ‘Body’ and ‘Winey’. ‘Body’ is very prevalent in the dataset and has a value greater than 0 for 97% of the whiskies in the database. ‘Body’ is also highly correlated with several other vaiables such as ‘smoky’ and ‘medicinal’, while also negatively correlated with the ‘Floral’ flavor.

‘Winey’ is a much less prevalent flavor, only being greater than 0 in about 65% of the whiskies examined here. It is however the most obvious distinguishing characteristic in cluster #4, and varies considerably between the other clusters.

The third most important variable is ‘Smoky’ which is one of the defining characteristics of cluster #1.

Conclusions

Flavor is as much an art as a science and takes a highly trained expert to be able to accurately gauge the 12 flavors examined in this dataset. Despite this, based on this analysis it does appear that whiskies can be classified into relevant clusters by their flavor profiles.

Much insight can be gained by examining the average values of each flavor for the different clusters but the variable importance output from the random forest algorithm proved to be very helpful in determining which flavors typically set different whiskies apart from each other.

References

1. Whiskey Classified: Choosing Single Malts By Flavor (https://www.amazon.com/dp/1862059136/ref=asc_df_1862059136?tag=shopz0d-20&ascsubtag=shopzilla_mp_1437-20&1548605092959157379551009030200800 5&creative=395261&creativeASIN=1862059136&linkCode=asn) Book from which dataset was originally sourced.↩
2. Wikipedia: Whiskey Tasting (https://en.wikipedia.org/wiki/Whisky_tasting) An introduction to the science/art of whiskey tasting.↩
3. Revolution Analytics Analysis of Same Dataset (2013) (https://blog.revolutionanalytics.com/2013/12/k-means-clustering-86-single-malt-scotch-whiskies.html)A similar analysis of the same dataset that uses some interesting R libraries such as Rmap.↩