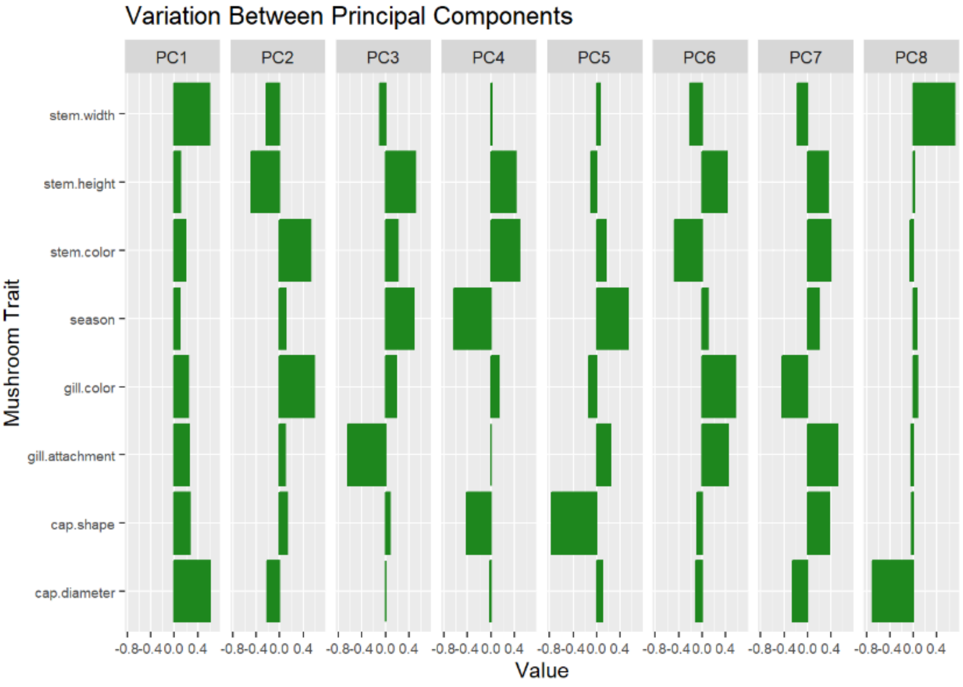
Mushroom Classification

Essential Questions

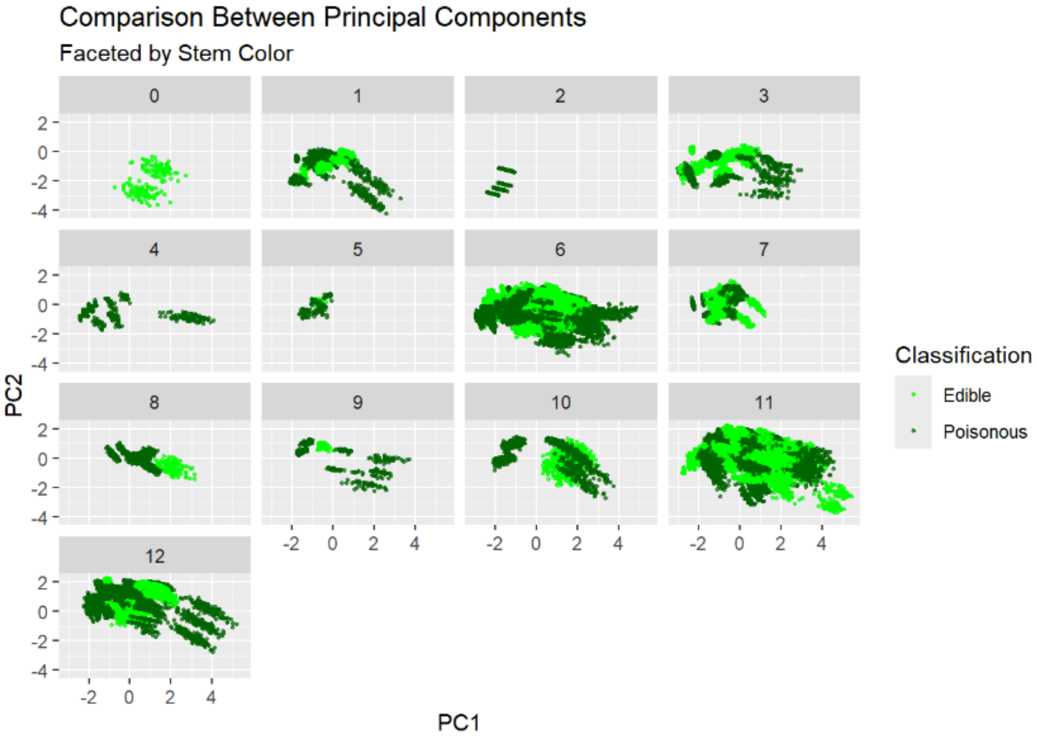
While interning this summer in a field work position, I spent a lot of time outside. Often my coworkers and I found mushrooms, and while they all looked the same to me, others would try to identify them. If they felt confident about its identity and its ability to be eaten, we would harvest them (if appropriate), but more often than not we were not able to tell if they were poisonous. I decided I wanted to look at a number of common mushroom traits to see if there was a way to distinguish between edible and poisonous mushrooms. Both visualizations serve to answer this question: what traits, if any, can be used to determine the status (edible or poisonous) of a mushroom?

Design Aspects

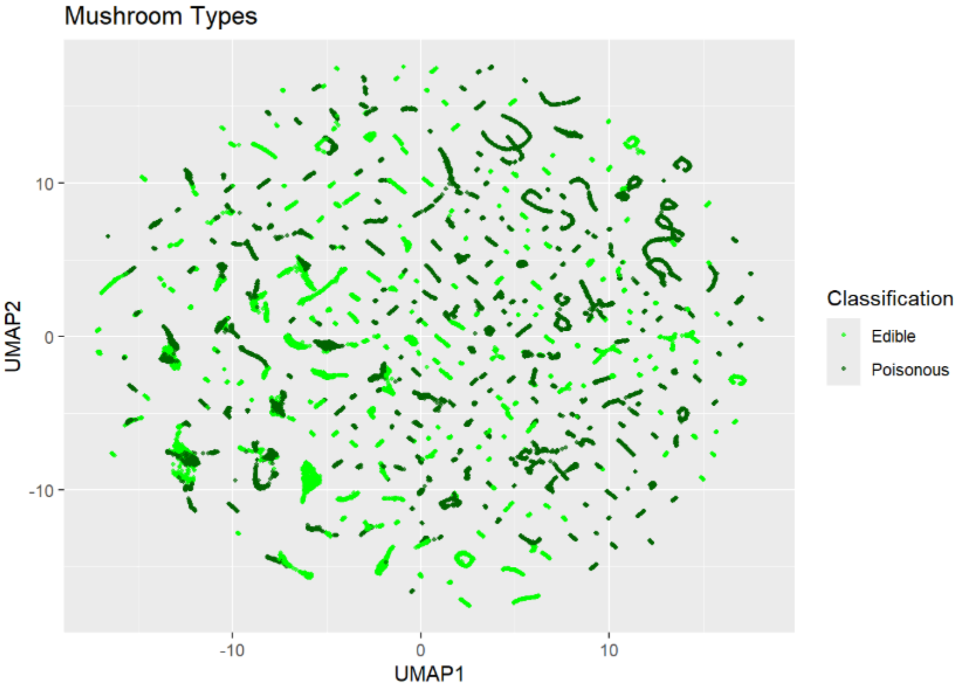
First, I did PCA to analyze the data. The first visualization is supplementary and looks at the variation between principal components, showing how the values of each trait differ. For example, the first principal component has a higher than average stem width and cap diameter, while the second principal component has a lower than average stem width and cap diameter, among other differences. While this visualization clearly shows the differences between principal components, it doesn't directly show the differences between the two types of mushrooms and will instead be used in conjunction with my second visualization.



For the second visualization, I chose to make a scatterplot showing the relationship between PC1 and PC2 values. I faceted the plot by stem color, because that is one of the most easily discernible factors when looking at mushrooms. Some facets show clear patterns, while others seem more randomized. One challenge with this visualization is that the dataset I found did not have documentation provided to explain the colors associated with the numeric encoding for the stem colors. So while I can see that all mushrooms with stem color ‘0’ are edible, that does not translate into real-world knowledge.



Next, I chose to visualize the data using UMAP.This will help me visualize if the mushrooms of one type are similar to each other, but it will not show me the attributes to look for when identifying whether mushrooms are poisonous or edible. As I’ll get into more detail later, some documentation for the numerically-encoded variables was not included, so it’s challenging to interpret the output for real life application. UMAP provides a visual representation of how closely mushrooms are related to each other, which can be useful to explain whether or not there are large differences between the groups, even if those exact differences cannot be quantified.



Key Findings

The key takeaway from these visualizations was that while there are similarities within mushroom types (poisonous or edible), there is not a clear way to identify them.

The PCA showed some patterns when comparing PC1 and PC2, but there was not a clear division of classification. Looking at the individual facets, some stem colors appear to be corresponding to exclusively poisonous or edible mushrooms, but again we are lacking the documentation to know what those stem colors are.

The UMAP visualization similarly showed chunks of related mushrooms within one classification, but an overwhelmingly randomized plot distribution. Mushrooms were similar to mushrooms of the same type, but there are also a lot of similarities between mushrooms of different types.

These findings relate pretty directly to my prior understanding. In my personal experience, edible mushrooms often have a poisonous look-alike, and there is no simple trick or metric that is used to tell them apart. The people who are able to identify them often use a combination of traits, including smells and textures, which were not included in the dataset. I was optimistic that I might find patterns in the data that would improve my mushroom identification skills, but unsurprisingly that did not happen.

Data Preparation

To create the visualizations, I used what I learned in class and my pre-existing knowledge of PCA. There was minimal data preparation involved. I was lucky enough to find a dataset with no missing values. I did change the variable for classification to have the words ‘poisonous’ and ‘edible’ as opposed to being encoded with zeros and ones. I was frustrated that many of the numeric features in the dataset did not come with documentation, therefore preventing me from making anything beyond general analysis on those points.