

**TDS3301 – DATA MINING**

**Lecturer: Dr. Ho Chiung Ching**

**Tutorial Section: TT01**

**Assignment #1: Exploratory Data Analysis**

|  |  |
| --- | --- |
| **Student ID** | **Name** |
| 1132701439 | Tay Tiong Soon |
| 1132700475 | Teo Yong Zhi |
| 1131122661 | Tazeek Bin Abdur Rakib |
| 1132701076 | Khairul Azmi bin Aminuddin |

**Dataset**

This dataset chosen for this assignment is related to classification of mushrooms. Though this dataset was originally to the UCI Machine Learning repository nearly 30 years ago, it is enjoying popularity nowadays. It consists of species of mushrooms, each identified as edible or not edible (poisonous).

There are about 22 attributes, excluding the class attribute. These attributes are categorized according to the parts of mushrooms: cap, gill, stalk, veil and ring. Some attributes corresponds to colors in various parts of the mushroom, odor smell from the mushrooms and the population along with the habitat of the mushrooms, as well as the spore color.

All the attributes are categorical. Hence, there are no numeric attributes present in this dataset.

**Possible Insights**

The possible insights that are in the dataset are:

* What features contribute most towards classification?
* What kinds of features can be clustered?
* Which features contribute unnecessary information?
* How are features correlated amongst each other?

Some examples of relationships between factors, as well as counts of factors, are shown below. The visualization was done using ggplot2 library:

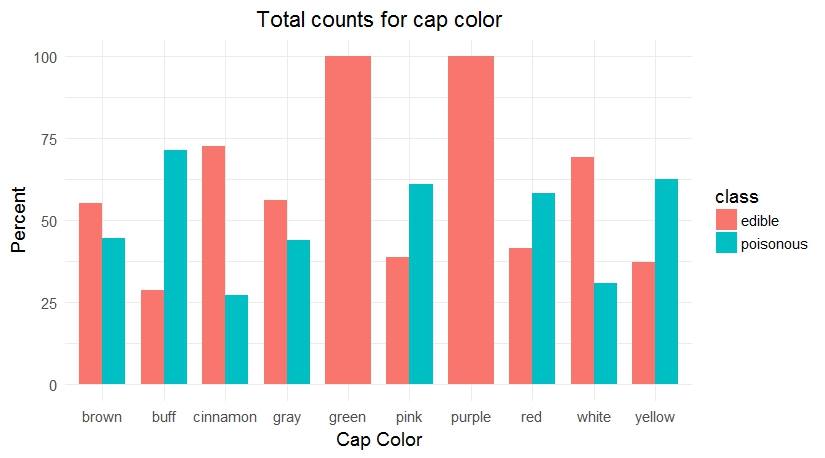


Figure 1: Edibility of mushrooms based on cap color

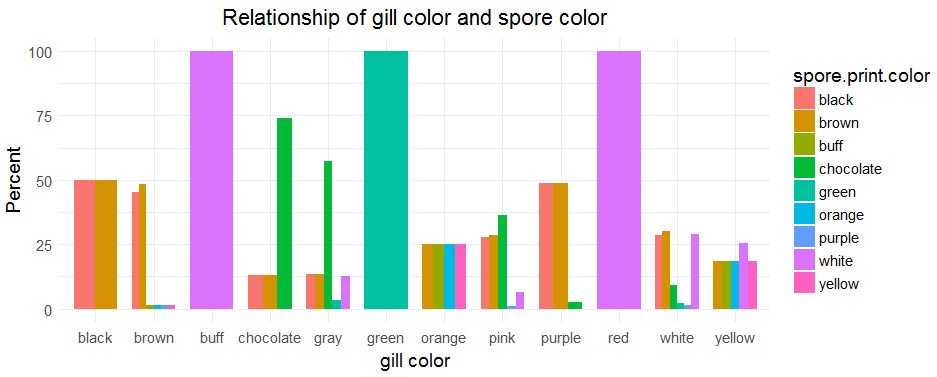


Figure 2: Correlation between spore color and gill color

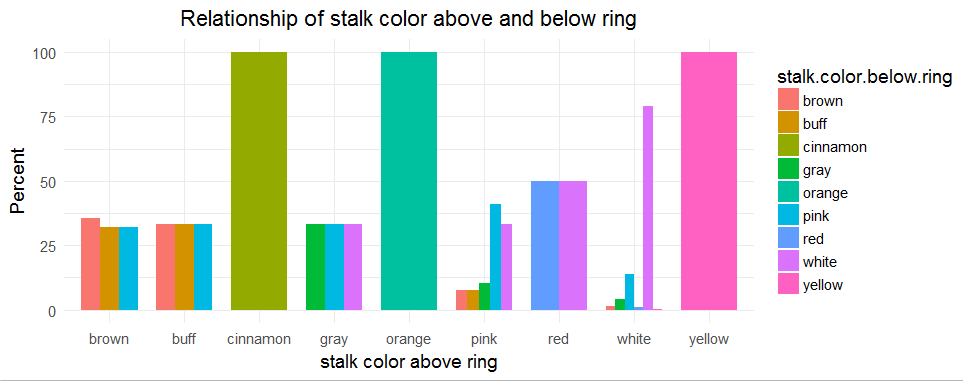


Figure 3: Correlation between stalk color (above and below mushroom ring)

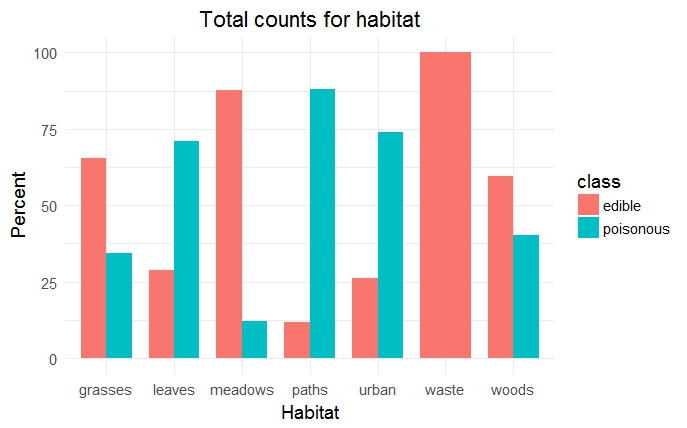


Figure 4: Edibility of mushrooms based on Habitat

**Data Mining Techniques**

From this dataset, the following techniques can be applied:

* **Classification:** Determine whether a mushroom is edible or non-edible. The dataset consists of an attribute named **class**, which is used to determine whether the mushroom can be eaten or not. Hence, this in turn leads to a **binary classification problem.**
* **Clustering:** Choose what attributes can be clustered into groups. For example, the properties of a mushroom’s stalk, which includes surface, root, stalk all have different properties and can lead to a variation of clusters. Some other possible clusters include taking the mushroom’s cap properties and gill into consideration.

**Issues**

The issues faced on this dataset were:

* **Missing values:** Some of the mushrooms’ stalk-root attribute were placed with a ‘?’
* **Confusion:** Attributes, mostly which describes color of mushrooms’ parts, are encoded in such a way, that there is difficulty in determining what exactly the attribute contains. For example, red corresponds to *e*, brown corresponds to *b*, and buff corresponds to *b*.

**Data Preprocessing**

The following preprocessing were done:

* **Missing values:** all ‘?’ were replaced with ‘unknown’.
* **Check for duplicates:** check whether the dataset contains any duplicate values or not.
* **Normalizing:** converting the attributes’ values into their respective full-form to reduce confusion. The values for the necessary normalization were obtained from the Kaggle website.

**Data Transformation**

The possible data transformation can be carried out:

* **Drop Column:** dropping the veil column, since it is 100% dominated by one value.
* **Label Encoding:** convert all unique strings to integers.
* **Correlation:** find features that contributes to the maximum information by using, for example, Pearson’s correlation.
* **Scaling:** scale the integer values into a suitable range, such as 0-1, for machine learning purposes.
* **Data Reduction:** reduce the dimensionality of the data using techniques such as Principal Component Analysis (PCA).