Project Proposal

Tofu-FC - Huiwen Wang, Rocky Zhang, Darrick Zhang

```
library(tidyverse)
library(tidymodels)
library(dbplyr)
# add other packages as needed

mushrooms <- read.csv2("~/project-Tofu-FC/data/mushrooms_dataset.csv")
```

Introduction

Subject Matter

Mushrooms are vital to the general wellness of the ecosystem, decomposing and recycling the nutrients in the soil. Mushrooms also provides a valuable food source full of nutrients for human beings and other important organisms. With this in mind, we are interested to see if there are any environmental factors and physical features of mushrooms that can help curious human avoid poisonous mushrooms that may grow in their yard. Ultimately the study of mushrooms is essential for safety and well-being.

"The ingestion of wild and potentially toxic mushrooms is common in the United States, with poison centers logging cases in the National Poison Data System (NPDS) for over 30 years. From 1999 to 2016, there were 133,700 reported cases of mushroom exposure, mostly unintentional and involving children under six years old. While the majority of cases resulted in no or minor harm, there were 704 instances of major harm and 52 fatalities, primarily due to cyclopeptide-producing mushrooms ingested unintentionally by older adults. Misidentification of edible mushroom species is a common cause of poisoning and may be preventable through education"(Brandenburg, William E, and Karlee J Ward. "Mushroom poisoning epidemiology in the United States." Mycologia vol. 110,4 (2018): 637-641. doi:10.1080/00275514.2018.1479561). Because of this research and that of the nature, accurate classification of mushrooms is crucial for preventing poisoning incidents. Many toxic mushrooms closely resemble edible varieties, making it easy for foragers to misidentify them. So our research will focus on **what physical features and environmental factors of**

mushrooms human can use to identify toxic/poisonous mushrooms in the wild. By conducting a research on how to distinguish between safe and dangerous species, we can mitigate the incidence of mushroom poisoning and ensure safer foraging practices.

Hypotheses:

1. Mushrooms in the wild with obvious physical features like white gills, white rings, red caps, or red stems tend to be poisonous. We want to explore this hypothesis because mushrooms with obvious physical traits are more likely to be spotted by animals, which would provide an evolutionary disadvantage unless they contain certain self-defense mechanisms, such as poison or toxins.
2. The habitat and season of the mushrooms would likely affect whether they're poisonous. We form this hypothesis because in specific habitats, mushrooms might face intense competition for resources such as nutrients, light, or space. Producing toxins can give them a competitive edge by inhibiting the growth of other fungi or microorganisms. In addition, the growing season can influence the chemical composition of mushrooms. Different temperature, humidity, and light during various seasons can affect the production of secondary metabolites, including toxins. For instance, some mushrooms might produce more toxins during wet seasons to fend off increased microbial activity.

Data description

The dataset we used can be found here: Mushroom Dataset

The data was curated on April 26, 1987, and submitted to the UCI by the National Audubon Society Field Guide. The National Audubon Society conducted extensive field research throughout North America, recording their observations on various aspects of mushrooms. Their research incorporate a wide range of physical characteristics, including size, shape, color, and texture of the mushrooms. Additionally, they documented environmental factors such as the type of habitat and seasonal variations. Importantly, the study also focused on the toxicity of the mushrooms, noting which species were poisonous. This comprehensive dataset provides valuable insights into the relationship between mushrooms and their environments, contributing significantly to the understanding of the factors influencing mushroom toxicity.

The dataset includes characteristics like cap diameter (quantitative), cap shape (qualitative), stem height and width (quantitative), type of stem root and color (qualitative), ring-type (qualitative), habitat (qualitative), and season (qualitative), and whether the mushroom is edible or poisonous (qualitative). The dataset contains a varieties of mushrooms from different families and species without any species or family dominating the dataset.

Exploratory data analysis

The data transformation process involves removing any observations with missing values in the explanatory variables of interest (view Analysis Approach section). However, since this dataset was previously utilized for competitions, any observations with missing values have already been excluded. As a group, we also determined that the dataset contains sufficient variables to avoid the need for joining it with another dataset. Additionally, given the limited number of quantitative variables, we opted to retain them in their original form rather than transforming them into categorical variables. Lastly, we needed to make all the variables that are quantitative into numerical data as R read them in as characters data type.

```
# Count the number of rows with at least one missing value
num_missing_rows <- sum(rowSums(is.na(mushrooms)) > 0)
cat("Number of rows with at least one missing value:", num_missing_rows, "\n")
```

```
Number of rows with at least one missing value: 0
```

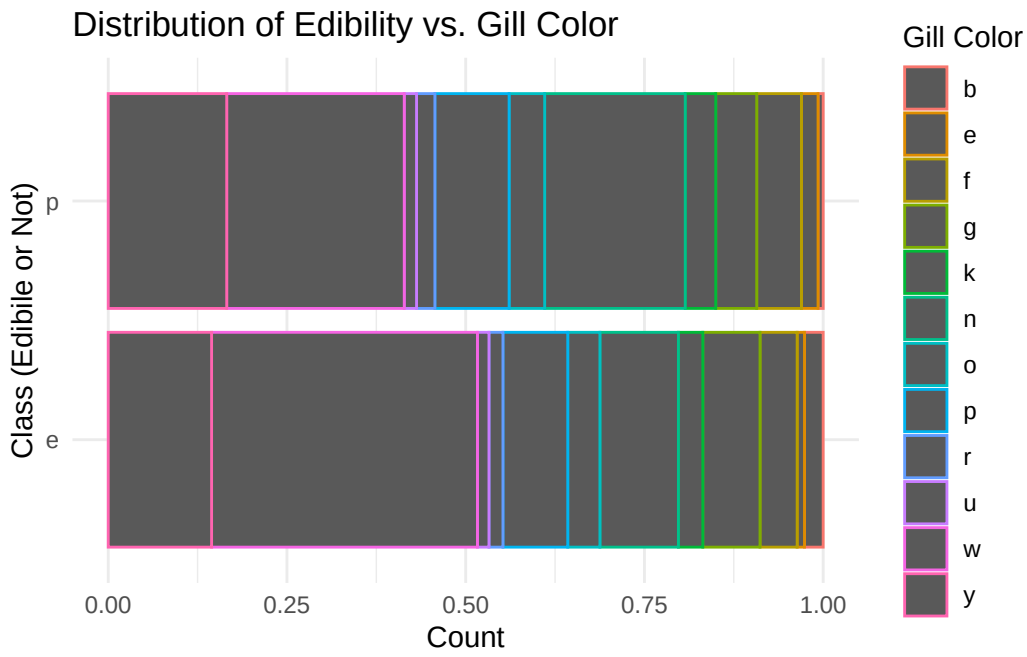
```
# Make Quantitative data numerical
mushrooms$cap.diameter <- as.numeric(mushrooms$cap.diameter)
mushrooms$stem.height <- as.numeric(mushrooms$stem.height)
mushrooms$stem.width <- as.numeric(mushrooms$stem.width)

# EDA Response (Categorical) Variable
ggplot(mushrooms, aes(x = class)) +
  geom_bar() +
  labs(
    title = "Distribution of Edibility/Classes of Mushrooms",
    x = "Class"
  ) +
  theme_minimal()
```


```
# Summary Statistics
mushrooms |>
  group_by(class) |>
  summarise(
    count = n(),
    percentage = (count / nrow(mushrooms)) * 100
  )
```

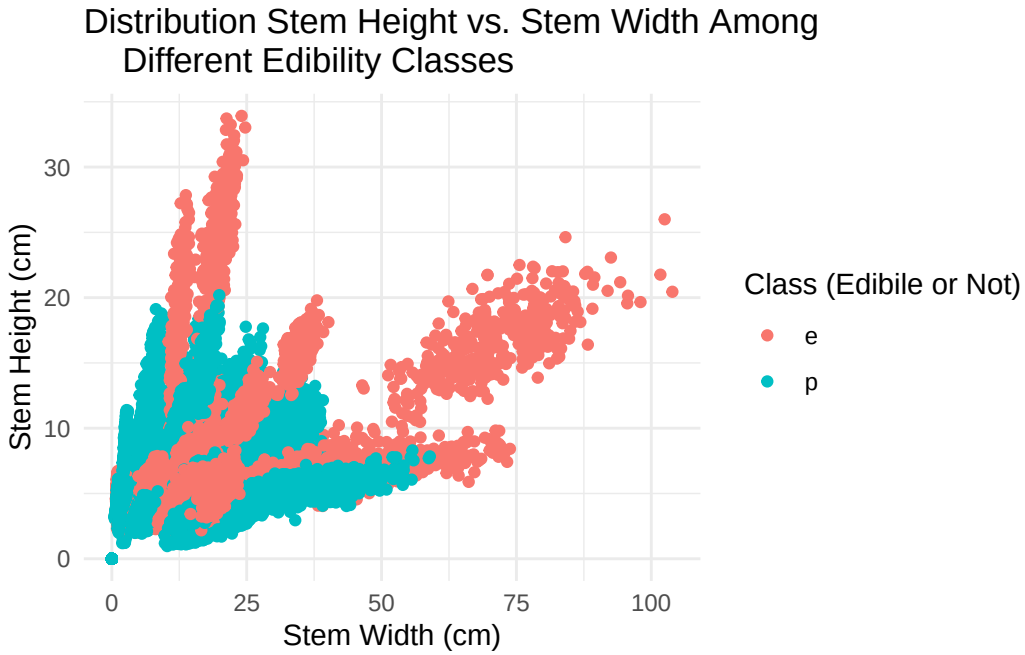
```
# A tibble: 2 x 3
  class count percentage
  <chr> <int>      <dbl>
1 e     27181       44.5
2 p     33888       55.5
```

```
# EDA Hypothesis 1
ggplot(mushrooms, aes(y = class, color = gill.color)) +
  geom_bar(position = "fill") +
  labs(
    title = "Distribution of Edibility vs. Gill Color",
    x = "Count",
    y = "Class (Edibile or Not)",
    color = "Gill Color"
```

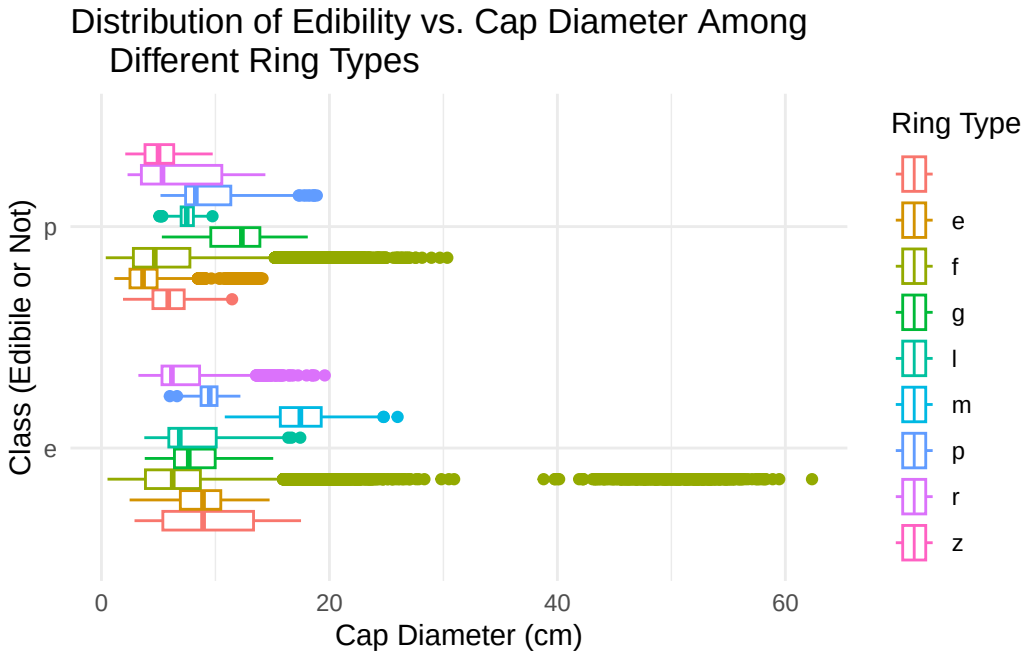
```
  ) +
  theme_minimal()
```



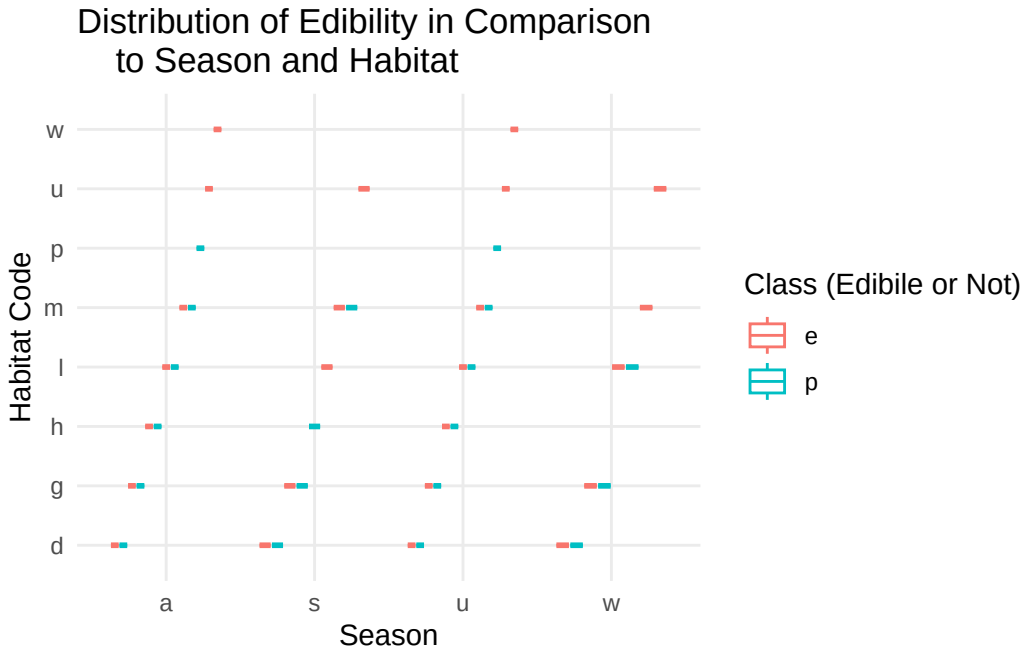

```
ggplot(mushrooms, aes(y = stem.height, x = stem.width, color = class)) +
  geom_point() +
  labs(
    title = "Distribution Stem Height vs. Stem Width Among
    Different Edibility Classes",
    y = "Stem Height (cm)",
    x = "Stem Width (cm)",
    color = "Class (Edibile or Not)"
  ) +
  theme_minimal()
```




```
ggplot(mushrooms, aes(y = class, x = cap.diameter, color = ring.type)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Edibility vs. Cap Diameter Among
    Different Ring Types",
    x = "Cap Diameter (cm)",
    y = "Class (Edibile or Not)",
    color = "Ring Type"
  ) +
  theme_minimal()
```




```
# EDA Hypothesis 2

ggplot(mushrooms, aes(y = habitat, x = season, color = class)) +
  geom_boxplot() +
  #facet_wrap(~season) +
  labs(
    title = "Distribution of Edibility in Comparison
    to Season and Habitat",
    x = "Season",
    y = "Habitat Code",
    color = "Class (Edibile or Not)"
  ) +
  theme_minimal()
```



The distribution of our response variables only shows two bars in the EDA as our response variable only has 2 categories (e and p). The distribution of the two categories are not too far apart as in their counts. There is a total of 27181 observations that falls into the class e and 33888 observations that falls into class p (6707 more observations). As for the actual percentages that the class makes up in the entire dataset, class e is among 44.51% of them while class p is the remaining 55.49%.

Note: The EDA for Hypothesis 1 and 2 are also shown out of curiosity!

Analysis approach

For the first hypothesis, we select variables representing apparent physical features like "cap.color", "gill.color", "ring.type", "veil.color", "cap.diameter" (quantitative), "stem.height" and "stem.width" (quantitative) as the explanatory variables.

For the second hypothesis, we'll focus on the environmental factors, using the "habitat" and "season" variable as the explanatory variable.

For both hypotheses, we'll use logistic regression because our response variable is a binary categorical variable ("classes" can only be "e" representing "edible" or "p" representing "poisonous").

Data dictionary

The data dictionary can be found here