**Data Proposal**

**Description of the dataset**

The ‘Mushroom’ dataset we have chosen includes descriptions of mushrooms from the Audubon Society Field Guide. It includes 61069 hypothetical mushrooms with caps based on 173 species (353 mushrooms per species). Each mushroom is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended (the latter class was combined with the poisonous class).

Of the 20 variables in the dataset, 17 are nominal and 3 are metrical. 8 of these columns contain significant amounts (>5% of observations) of null values. Several of the columns are highly correlated with each other, but multicollinearity overall appears to be generally low (see correlation heat map at bottom of document).

**Data collection**

The dataset was submitted by Dennis Wagner to the UCI Machine Learning Repository on September 5, 2020. The mushroom species were drawn from the 1999 book ‘Mushrooms and Toadstools’ by Patrick Harding.

This dataset is simulated using a Python script from an original dataset of mushroom species containing 173 observations. This could affect the data because we’re estimating nearly 70000 observations based on less than 200 actual observations. This could exaggerate trends that aren’t real.

**Our question**

We will use various statistical and/or supervised machine learning algorithms to predict mushroom edibility based on the collected features found within the mushroom dataset.

