I am Lize Mostert, and I will be explaining the process of how we introduced missing values to our dataset as well as the methods applied during the imputation of these missing values.

The first step was importing a clean and complete dataset that was used in the research paper we are comparing our results to. It is worth noting that during the initial data exploration, we did observe a skewed class distribution in the target feature. Following the data importation and exploration the dataset was then split into a 70:30 ratio of training to testing subsets.

I then used the training dataset and introduced missing values randomly using a loop that iterated over every feature in the dataset. The missing values were introduced in three different proportions: 10%, 40% and 70% for each feature to create three different datasets with missing values from the training dataset.

The following step involved imputing the missing values that were introduced in the previous step. Two methods were applied namely mode imputation and Naïve Bayes imputation. Because our data consisted of only categorical features, the mode of each feature was determined and used to fill in the missing values. For the Naïve Bayes imputation, I built and trained the Naïve Bayes model for each feature, and assigned the predicted values from the model to the missing values. From the imputation methods, two separate data frames were created; the mode\_imputation and nb\_imputation for the respective proportions of missing values.

The final step was to map the numerical values assigned to the categorical variables by the Naïve Bayes model back to the original classes for each feature and calculating the NA counts for all the data frames created to ensure that missing values were all imputed.

At the end of this process, we were left with 8 datasets. A testing set consisting of 30% of the original dataset with no missing values, a training set consisting of 70% of the original dataset with no missing values, 3 training sets imputed with the mode for 10, 40 and 70% missing values that were introduced, and 3 training sets imputed via the Naïve Bayes model for the 10, 40 and 70% missing values that were introduced.