**1. Exploratory Data Analysis (4000 5% / 6000 3%) - Examine a subset of features in the dataset both separately and in pairs. This may involve cleaning the dataset, for example removing missing values, applying transformations such as log scale and/or taking subsets. You have complete freedom in the selection of features to explore and any necessary filtering. Make sure to use suitable plots to examine variable distributions as well as pairwise relationships between variables. Consider how the presence of outliers affects the plots and consider removing them if they have a significant effect. Explain what you learned about the dataset in 5-6 sentences.**

For this dataset, I focused on several morphological attributes of bean samples, including area, perimeter, diameter, and related geometric measurements. I began by generating boxplots for each variable to identify potential outliers and assess the overall distribution of values. Although some points appeared to be outliers initially, further inspection revealed that these values were characteristic of a particular bean class with naturally larger measurements rather than true outliers. To better understand this distinction, I examined the distribution of each attribute within individual bean types, which clarified how class-level differences influenced the plots with all of the data. This analysis highlighted the importance of considering class-specific variation before removing data points. Overall, the EDA provided a clearer understanding of how the features differ among bean categories and how those differences shape the dataset as a whole.

**2. Predictive Modeling (5%) - Decide on a problem to solve by developing predictive models using the data. This could be the same problem for which the dataset was intended or a different problem that you believe can be addressed with this dataset. The solution could involve predicting a continuous variable (regression) or a categorical variable (classification). Consider what you learned about the variables during exploratory analysis to decide which features should be used as inputs to the models. The response variable (output of the model) is usually clear and is dictated by the problem being addressed. Train and evaluate two models and compare their results. The two models should utilize different algorithms (e.g. kNN and Random Forest) but the same set of input variables. Classification models should be evaluated and compared using a confusion matrix and Precision/Recall measurements. Regression models should be evaluated using Mean Squared Error (MSE). Explain how models performed and comment on the suitability of the dataset to solve the problem you chose in 5-6 sentences.**

For the predictive modeling component, I focused on classifying each bean into its correct category using selected geometric features. Based on the insights from my exploratory analysis, I chose area, perimeter, major axis length, and minor axis length as the input variables for all models to ensure consistency. I trained two classification models, KNN and Random Forest, using the same feature set. The KNN model performed strongly, producing high accuracy along with strong precision and recall across most classes, indicating that these four features are effective predictors of bean type.

```
> confusion
          Actual
Predicted  BARBUNYA BOMBAY CALI DERMASON HOROZ SEKER SIRA
  BARBUNYA      226      0   52        0    16     4    6
  BOMBAY          0    148    0        0     0     0    0
  CALI          133      1  412        0    20     1    2
  DERMASON        0      0    0      951    12    46   95
  HOROZ          31      0   20        1   505     1   19
  SEKER           2      0    1       23     0   487   17
  SIRA           22      0    4       87    46    55  638
> accuracy
[1] 0.8244368
```

```
> knn.precision
 BARBUNYA    BOMBAY      CALI  DERMASON     HOROZ     SEKER      SIRA
0.5458937 0.9932886 0.8425358 0.8954802 0.8430718 0.8198653 0.8211068
> knn.recall
 BARBUNYA    BOMBAY      CALI  DERMASON     HOROZ     SEKER      SIRA
0.7434211 1.0000000 0.7240773 0.8614130 0.8752166 0.9188679 0.7488263
```

The Random Forest model performed even better, achieving higher classification accuracy and more consistent precision/recall metrics, likely due to its ability to capture nonlinear relationships and interactions among the features.

```
> rf.confusion
          Actual
Predicted  BARBUNYA BOMBAY CALI DERMASON HOROZ SEKER SIRA
  BARBUNYA      371      1   21        0     3     2    4
  BOMBAY          0    148    0        0     0     0    0
  CALI           31      0  454        0    17     0    4
  DERMASON        0      0    0      983     3    11   82
  HOROZ           5      0    7        8   559     0   10
  SEKER           3      0    0       11     0   566   16
  SIRA            4      0    7       60    17    15  661
> rf.accuracy
[1] 0.9162586
```

```
> rf.precision
 BARBUNYA    BOMBAY      CALI  DERMASON     HOROZ     SEKER      SIRA
0.8961353 0.9932886 0.9284254 0.9256121 0.9332220 0.9528620 0.8507079
> rf.recall
 BARBUNYA    BOMBAY      CALI  DERMASON     HOROZ     SEKER      SIRA
0.9228856 1.0000000 0.8972332 0.9110287 0.9490662 0.9496644 0.8651832
```

Overall, this dataset is highly suitable for the classification task, as the features are both descriptive and strongly correlated with bean type, enabling reliable predictive performance.

**Note:**
Code for this assignment can be found in my GitHub Repo:
https://github.com/brooke-kinsey/DataScience/blob/main/assignment5/assignment5.R