Introduction

* Intent of the application
  + Explore relationships within the Iris Dataset utilizing clustering techniques. This application will also compare data after which Principle Component Analysis is performed to determine if it is beneficial in this case.
* Dataset to be used, including source
  + Iris dataset from sklearn database
* Use case
  + Students or others wishing to learn more about the relationships and patterns of the features of an iris, and how they relate to which class each iris belongs to, can utilize this application to visualize and numerically understand those relationships in terms of clustering. In this case, there is no option for the user to decide which clustering techniques are used, it is already embedded in the program.

Dataset Analysis

* Define variables
  + Sepal length, sepal width, petal length, petal width
* Define labels
  + Setosa, versicolor, virginica

Proposed Libraries

* Libraries
  + Sklearn
  + Pandas
  + pandas\_profiling
  + Numpy
  + seaborn
* Library source
  + Scikit-learn.org
  + Pandas.pydata.org
  + <https://github.com/ydataai/pandas-profiling>
  + Numpy.org
  + Seaborn.pydata.org

Proposed Solution

* Functional description
  + This application continues to explore the Iris dataset through the use of different clustering techniques and Principal Component Analysis. It will produce visualizations of the created clusters for better understanding and will also explore how PCA improves or diminishes the results of clustering in this case.

Proposed Outputs

* Define application outputs
  + Features resulting from Principal Component Analysis
  + Conclusions on results of all analysis

Proposed Visualization

* Define visualization of outputs
  + Cluster visualizations
  + Hierarchical visualizations

Conclusions

* Analysis of results
  + The expected conclusion from this application will be the creation of clusters which match with the labels of the iris flowers. It will be able to showcase what clustering techniques provide the best insight to this dataset. However, the clusters created with the PCA features may be less accurate, as reducing the dimensions of an already low dimension dataset could result in losing important features that help with distinguishing between data points. The Virginica and Versicolor groups will likely have the greatest overlap, as they overlapped most in the EDA.