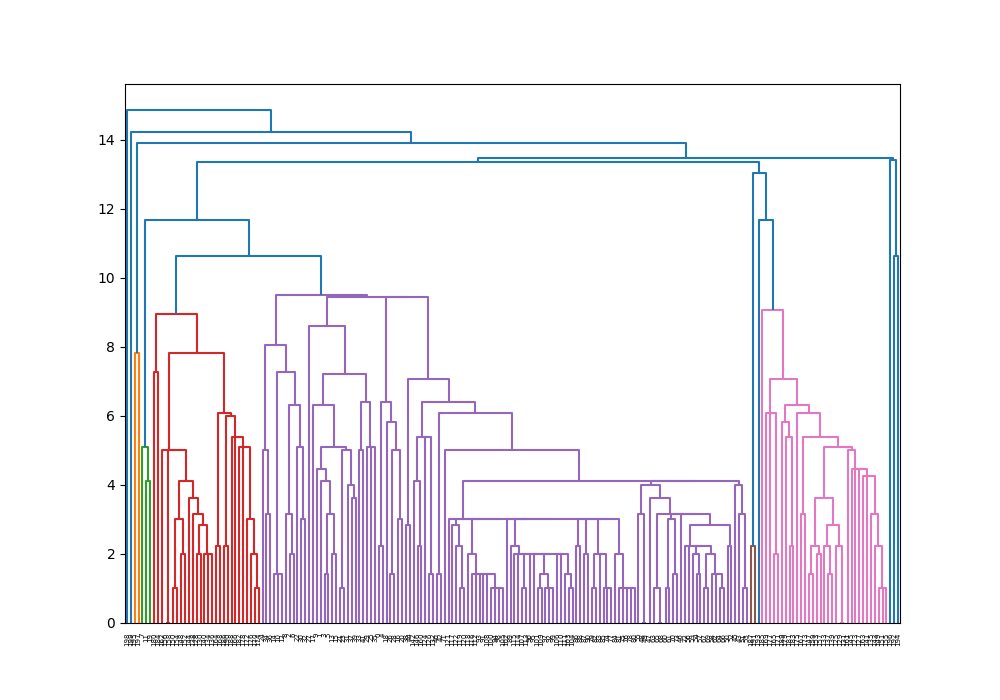
Lab1 Answers

# Part 1

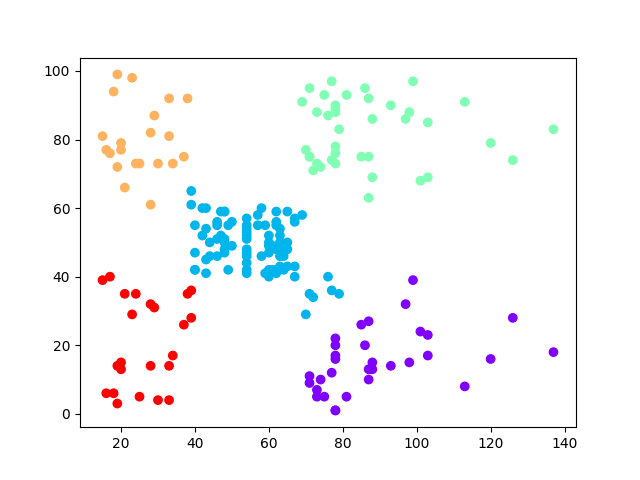
1. **What are the relevant features of the Titanic dataset? Why are they relevant?**  
   Depending on the use case, different features are relevant. For the use case in the tutorial, “dead or not”, the relevant features are: Pclass, SibSp, Sex, Parch, Age.
2. **Can you find a parameter configuration to get a validation score greater than 62%?**No, we did not manage to find such configuration.
3. **What are the advantages/disadvantages of K-Means clustering?**Advantages is that it is easy to implement, scales to large datasets, guarantees convergence and is easy to adapt to new datasets.  
   One disadvantage is that it takes good knowledge of the dataset (because it is a supervised method). Another on is that the method is sensitive to outliers and the order of the data.
4. **How can you address the weaknesses?**One big weakness of K-means is that it does not handle missing values well. To solve for this, we filled the empty columns with the mean values. Additionally, for the model to perform better we selected only the relevant features.

# Part 2

1. **How many clusters do you have? Explain your answer.**  
   The number of clusters is depending on where you draw your horizontal line. Looking at the dendrogram makes it hard to tell the number of clusters. However, the different colors would suggest six different clusters.



1. **Plot the clusters to see how actually the data has been clustered.**



1. **What can you conclude by looking at the plot?**  
   That it is five obvious clusters, this was hard to see with the dendogram.

# Part 3

1. **Can you choose n components=2? Can you think of some method to test this?**

Yes, the first two components are sufficient to separate the data. We tried with n=5 and used the third and forth component, which clearly showed that the data was hard to separate.

1. **Create the scatter plot of the third principal component (that is, you combine the third principal component with the first and then the second principal component). What can you see with the plot? What is the difference?**  
   The data is easier to separate with the principal components with highest addition to the total variance (first and second contribute to more than ¾ of the variance).
2. **Can you tell which feature contribute more towards the 1st PC?**   
   Mean concave points! (We checked the color with an eyedropper)