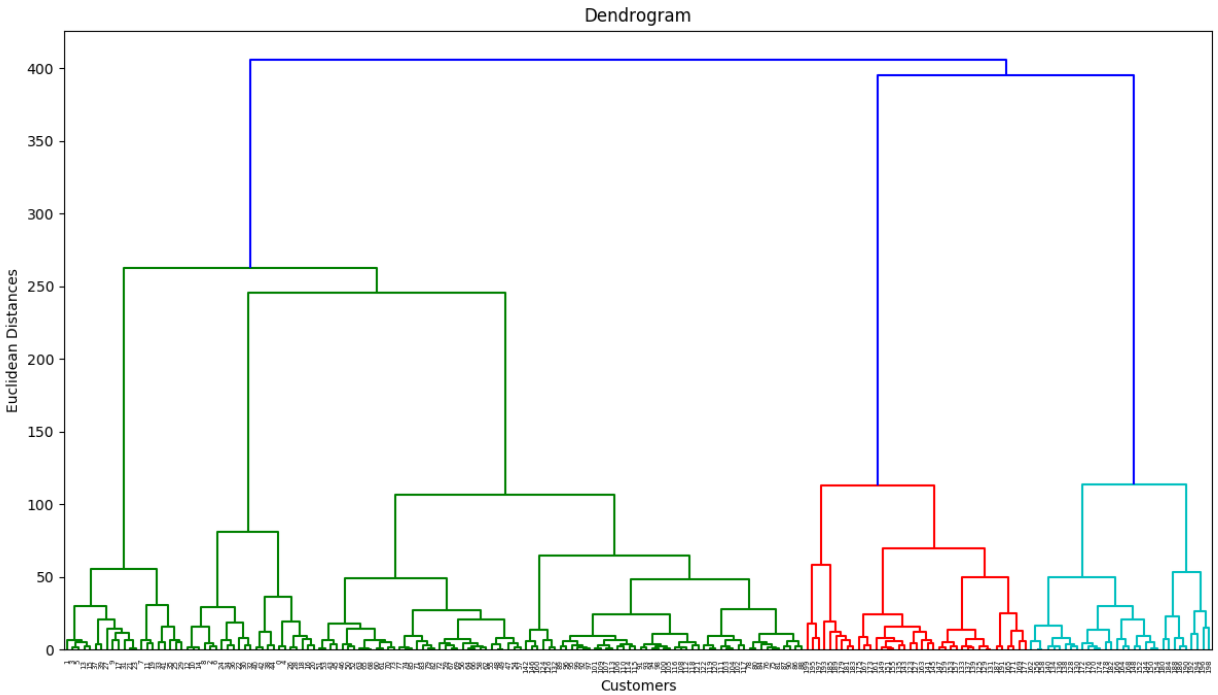
Clustering Findings

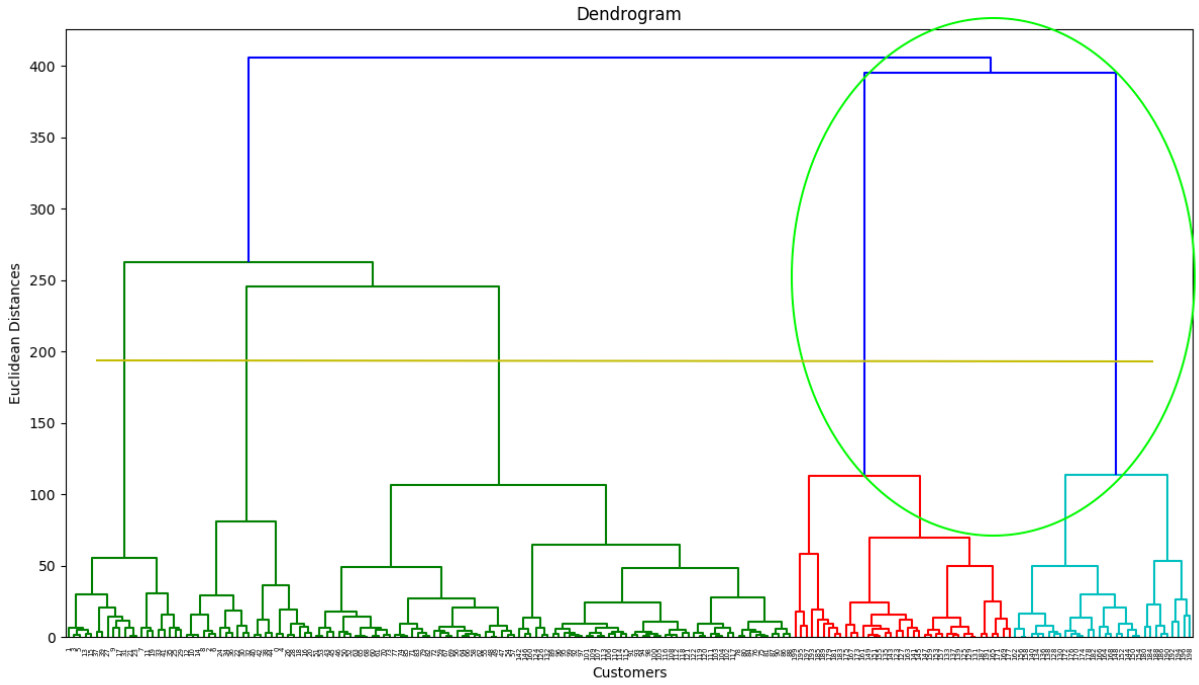
We are given a dataset (*Mall_Customers.csv*) that contains the information of customers. The dataset shows their gender, age, annual income, and spending score which is based upon how much they have spent. Our goal is to separate the clients into groups based upon their annual income and spending score.

Hierarchical Clustering

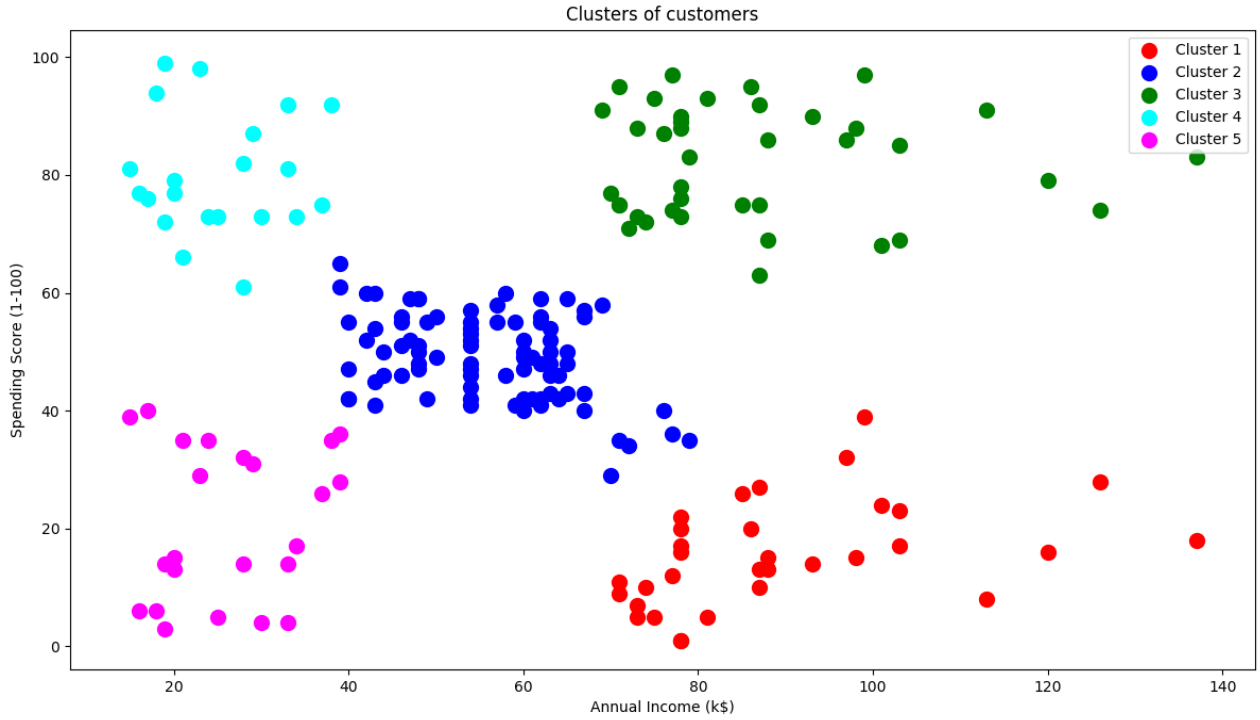
We built a dendrogram in order to determine how many clusters we should split the data into.



To determine the properly number of clusters find the largest vertical distance without crossing any horizontal lines (lime green circle). Draw a horizontal line (gold line) and count the number of vertical lines cross it.



In this case the number of clusters we should use is 5. After running the algorithm, we receive the following results.



For a slightly different looking graph see preview below (note this graph was created in R rather than Python so the results are expected to be slightly different).