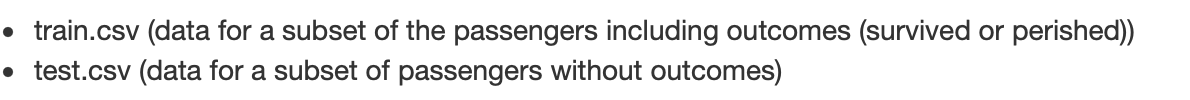
**Lab 1**

**Part 1**

Info:

 *1. What are the relevant features of the Titanic dataset. Why are they relevant?*- The relevant features are features that have an impact on the survival status of the passengers, such as *Pclass (*passenger Class, 1-3), *Age*, *SibSp (*number of siblings/spouses aboard), and *Parch (*number of parents/children aboard). Features as *Name*, *Ticket*, *Cabin* and *Embarked* do not tell us anything about the survival status.

To only have a model with significant features efficient the modelling and the training of the model can happen in much lesser time.   
 *2. Can you find a parameter configuration to get a validation score greater than 62%?*

- The features contain different ranges of values, for example sex only have 2 different values and age can vary from 0-100. A small change in a feature does not affect the other features.

We therefore must *scale* to a uniform value range across all the features.   
 *3. What are the advantages/disadvantages of K-Means clustering?*

- One of the disadvantages with K-Means clustering is that it requires the developer to pre-specify the number of clusters (k). In the Titanic dataset we had some knowledge available that told us the number of people who survived in the shipwreck, but this is not always the case with real world datasets.   
K-Means is also sensitive to outliers which is a term for a divergent value which is a disadvantage. An additional disadvantage is that different result of K-Means can occur if the user changes the ordering of the data.

On advantage with the very popular clustering technique K-Means is that the algorithm will help the user to tackle unlabeled datasets (for example the dataset that do not have any class-labels) and draw their own inferences (slutsats) from the with ease.

K-Means can also scale large data sets and is relatively simple to implement.

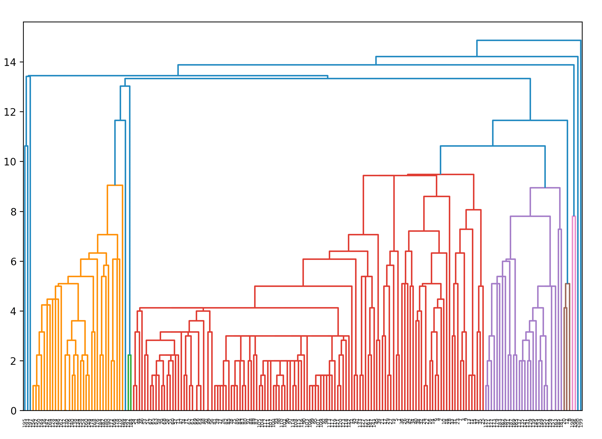
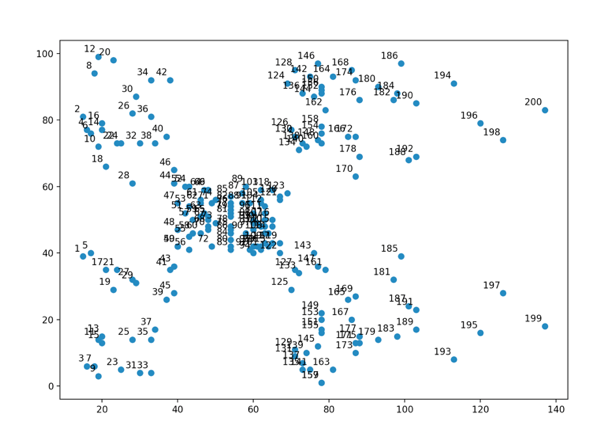
*4. How can you address the weaknesses?*

- One of the disadvantages with K-Means clustering was that it was required to pre-specify the number of clusters. Hierarchical clustering is an alternative approach that does not require a particular choice of clusters.   
  
Also, one weakness was that K-Means does not support missing values in the data. Therefore, we need to handle the missing values present in the data. This can be done by removing rows with missing values or by imputing missing values. The second one is preferred since removing rows with missing values can cause insufficiency in the data.

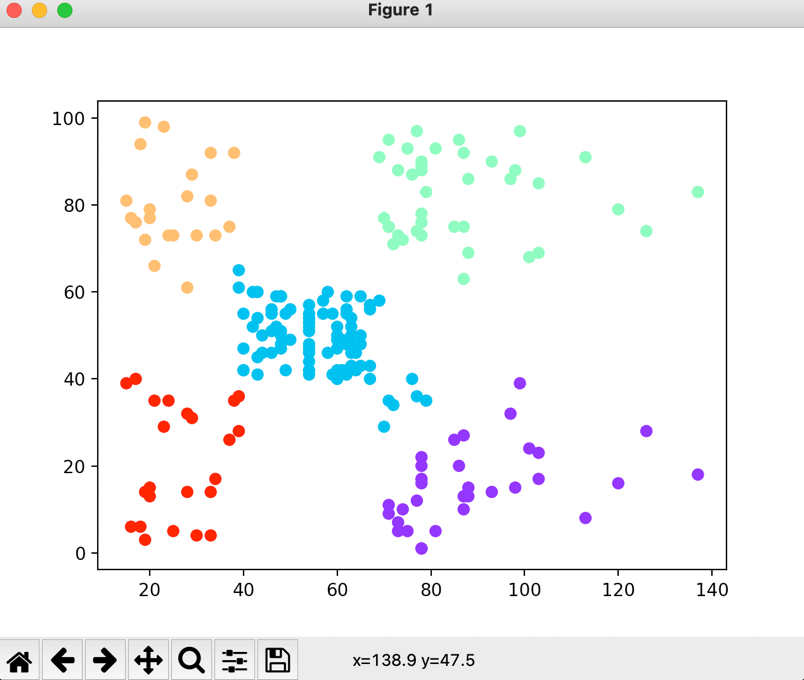
The imputation can be performed by the mean value for the column with the help of fillna() which is provided by Pandas.

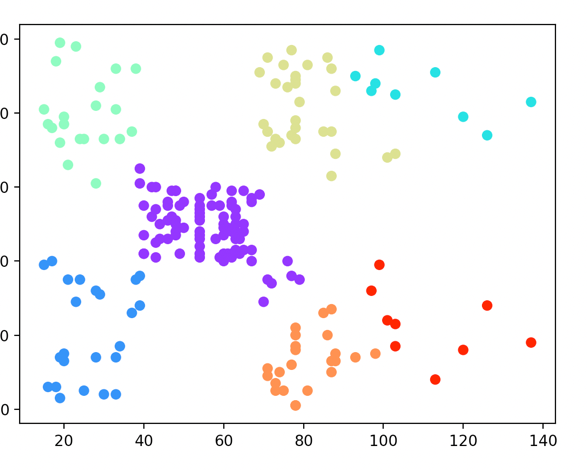
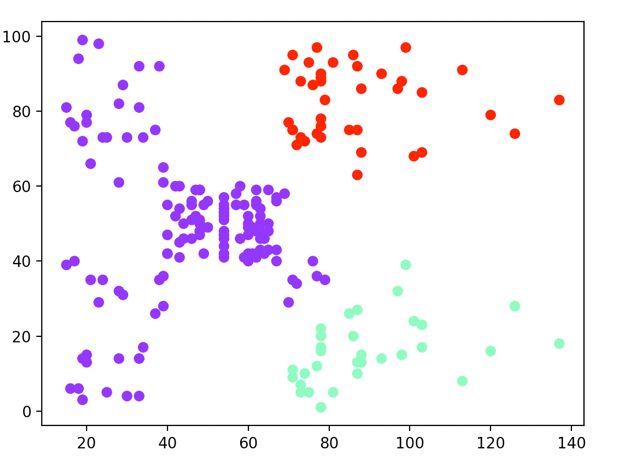
**Part 2**

***1.*** *How many clusters do you have? Explain your answer.*   
  
- We think that it seems to be five clusters since it´s in a formation as the five on a dice.

**

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

- It seems to be five!  
 **  
  
If we change the number of clusters in the plot, for example to 3 and 7, we got more scattered clusters which is in



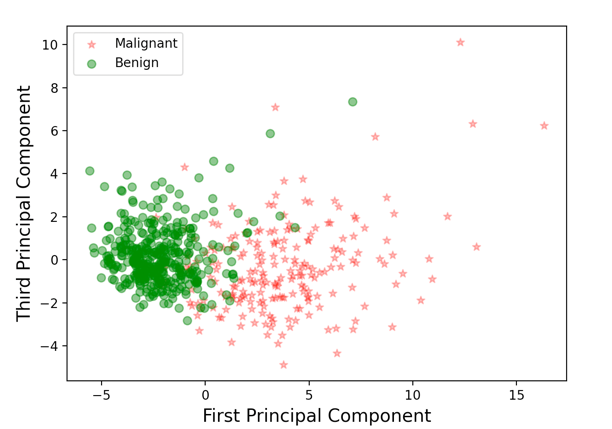
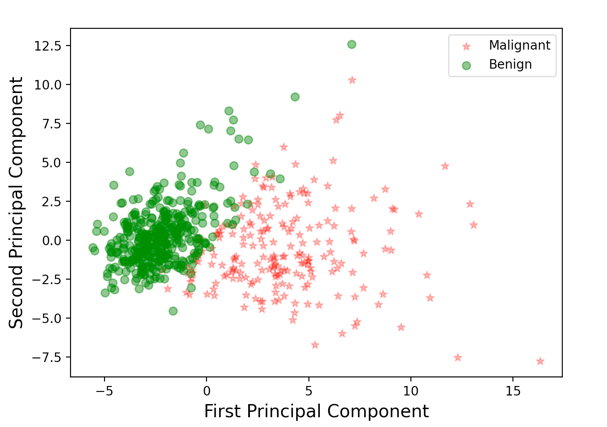
*3. What can you conclude by looking at the plot?*

- By looking at the plot we can conclude that the costumers that has an income around 40-70 k has the middle spending score, while the low- and high-income takers either spends a bit less than average or much more than average.

**Part 3**

*1. Can you choose n components=2? Can you think of some method to test this? (question at page 6)*

- Yes we can and we also get a better total variance, 99%, compared to earlier when we had 3 components and a total variance of 87%.

2. 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?  
(question at page 7)   


*-* The first plot is more scattered in both malignant and benign.

*3. Can you tell which feature contribute more towards the 1st PC? (question at page 8)*

En bild som visar text

Automatiskt genererad beskrivning  
- Mean concave points contribute more towards the 1 PC.

**Part 1:   
K-means** – centroid-based clustering. Centroid is a data point (imaginary och real) at the center of a cluster.   
K-means uses Manhattan distance to measure the distance between centroid points and the other data points and cluster them.   
  
**Part 2:**  
**Hierarchical clustering** – cluster unlabeled data points. Similar to K-means (clusters groups together with the data points with similar characteristics). The result can be similar with hierarchical and k-means.   
Distance can be Euclidean or Manhattan. You can choose whatever pivot point you want (closest, farthest, centroids or combinations).

**Part 3:   
PCA -   
En bild som visar text

Automatiskt genererad beskrivning**  
You can use PCA when deciding which features to choose so that the model is safe from overfitting. Also PCA is a method that reduces the dimension of the feature space in such a way that new variables are orthogonal to each other (for example they are not independent or not correlated).   
In PCA you probably need to scale the data (*StandardScaler*) such that each feature has unit variance. This is necessary because fitting algorithms highly depend on the scaling of the features.