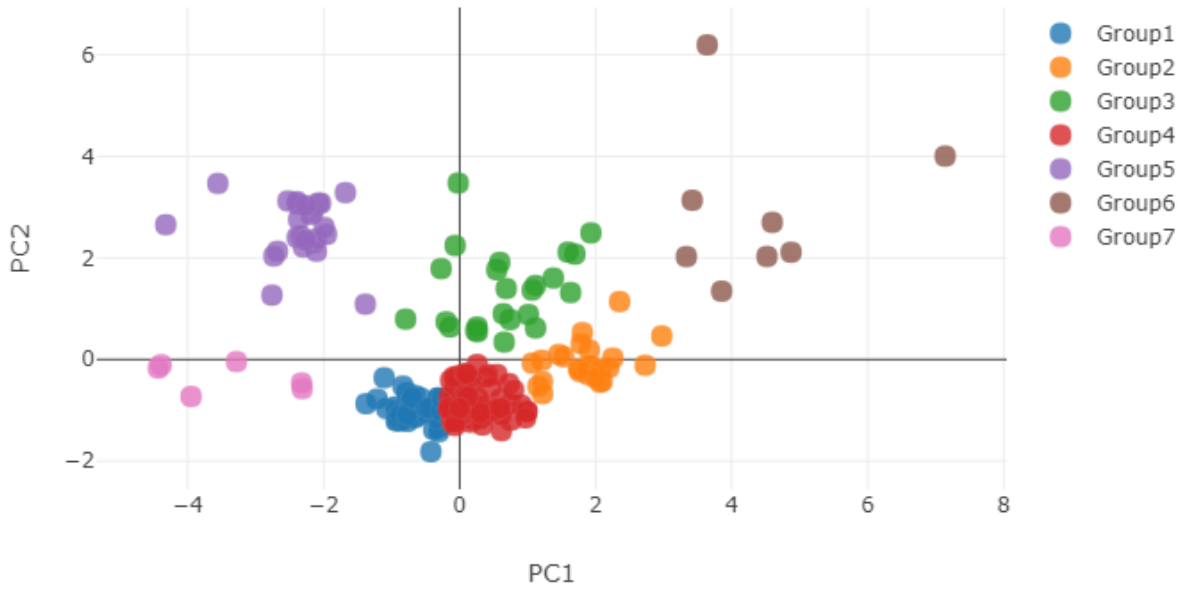
IDEAS WORKSHOP SUMMER 2017
WEEK 3

UTHAIPON TANTIPONGPIPAT

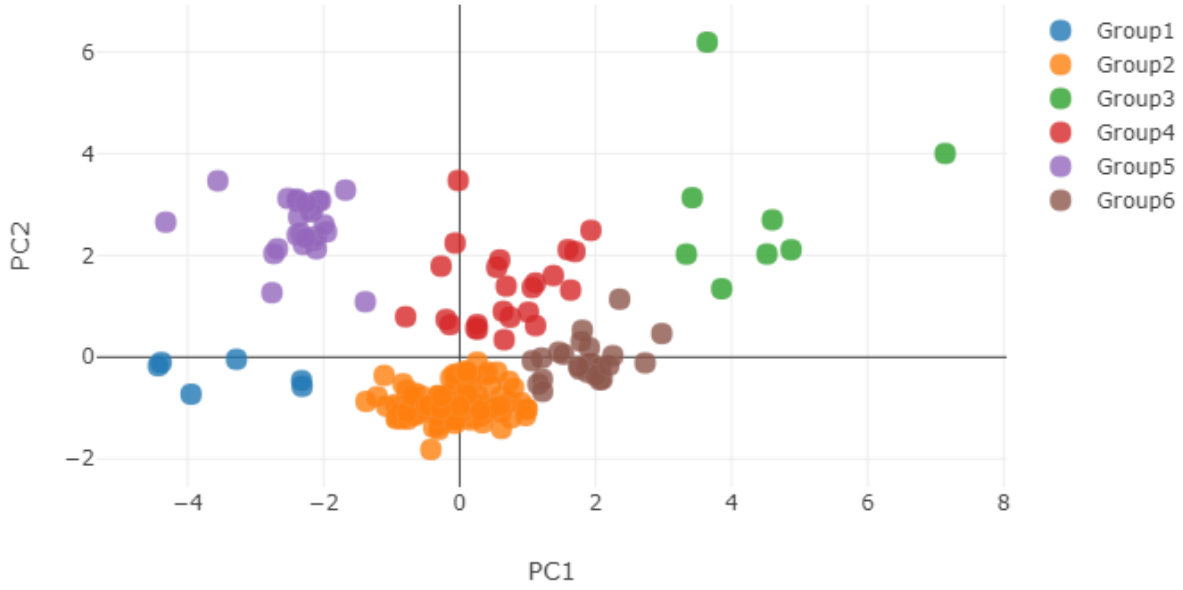
The tutorial gives a general idea of what clustering is, and the function to call clustering

```
kmeans = KMeans(n_clusters=d).fit(X)
```

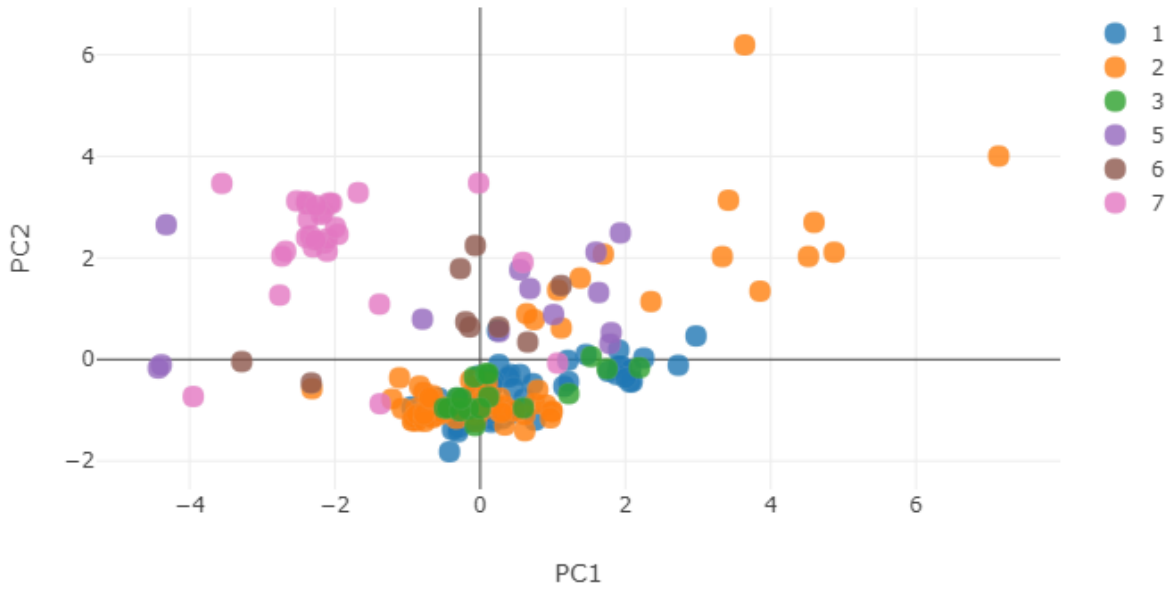
But the tutorial does not have the exact code to explain what data type of object X is, and how to plot the clusters. There is small incompatibility between code in week 2 and week 3, so I need to learn Scatter and a little bit about how to use dataFrame just to fix a small error. Just need to google quickly. I also hope that the tutorial will give access to the code of some measure of how good the clustering is, and so we can plot a graph "Elbow Point Example" on our data set as well. Without the exact measure to know how many clusters I should use, I tried from 2 to 10 clusters. The original data is to classify 7 types of glasses, one of which is not present in the data set. In my opinion, the appropriate number of clusters is 6. It's not because there are originally 6 classes of data, as the plot of PCA data shows that many different types of data are clustered together, so these 6 clusters do not correspond to original 6 glass types. The reason I choose 6 clusters is because after the number of clusters is at least 7, there are two clusters in the middle that have many points too close to one another that I think should be considered as the same cluster. In the picture which shows when the number of cluster is 7 below, those two clusters are red and blue.



For 6 clusters (picture below), only purple and green clusters seem to correspond to any meaningful glass types. Other clusters include a significant portion of at least 3 glass types.



For reference, the picture of PCA data of all glass types is present below:



The link to my code is: https://github.com/uthaipon/SkillsWorkshop2017/tree/master/Week03