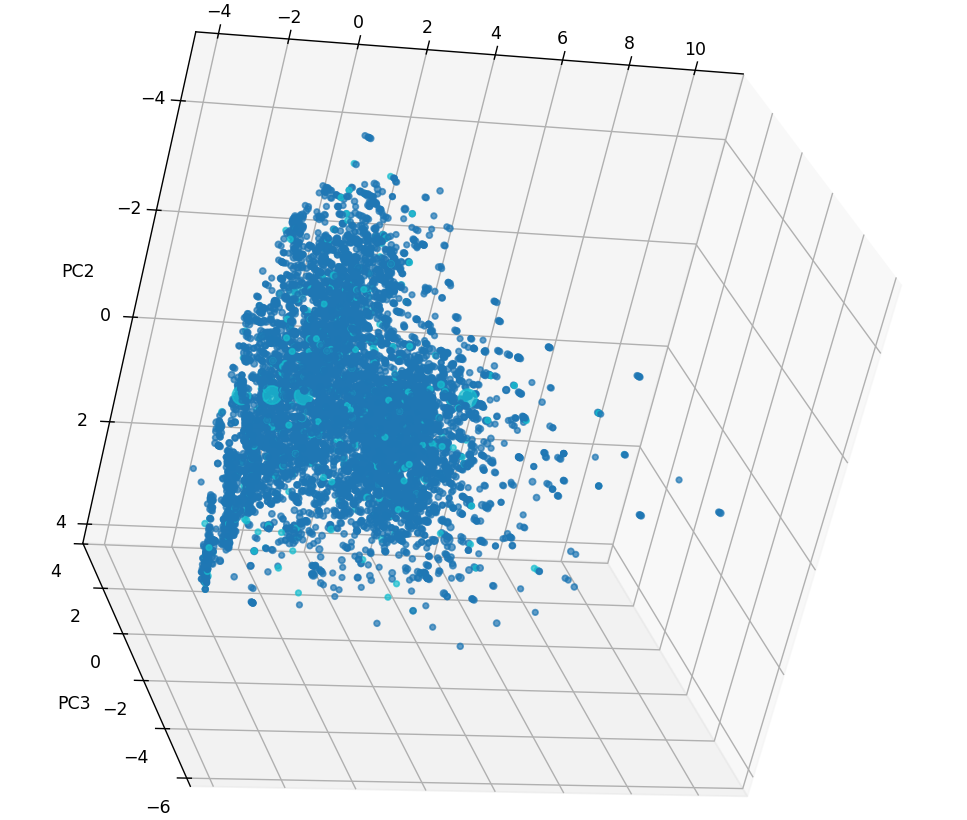
**IML – Hackathon**

**Part 3**

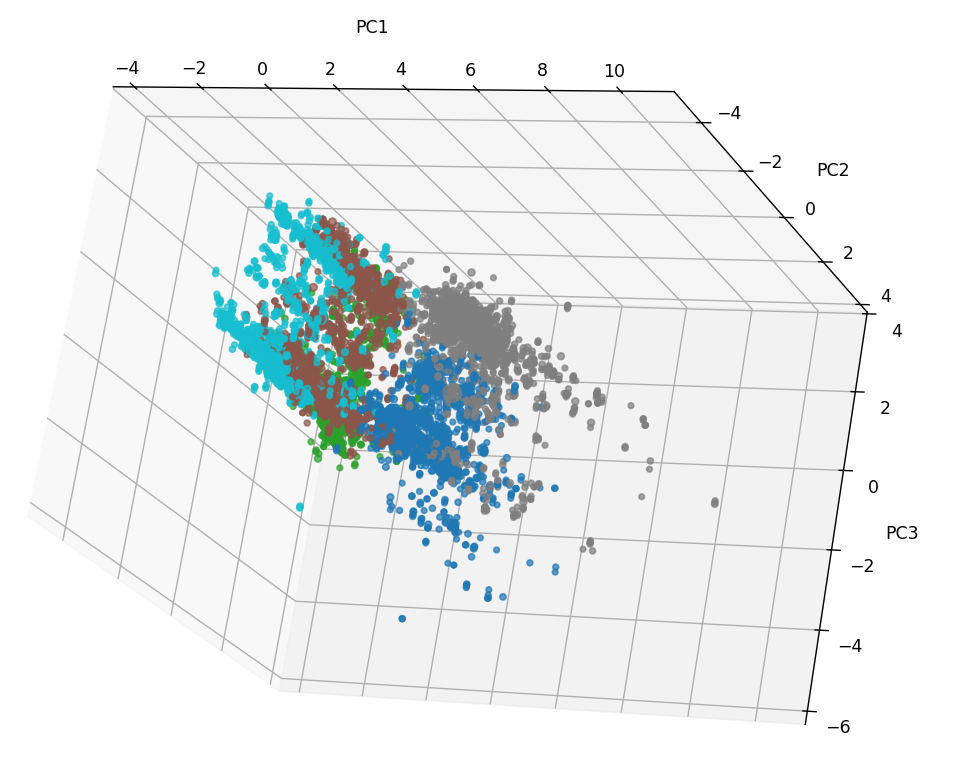
* We created a 3D PCA graph, where each dot has a size proportional to the size of the tumor, and colored according to the existence of metastases (a base size was added so points representing samples without tumors are still visible).

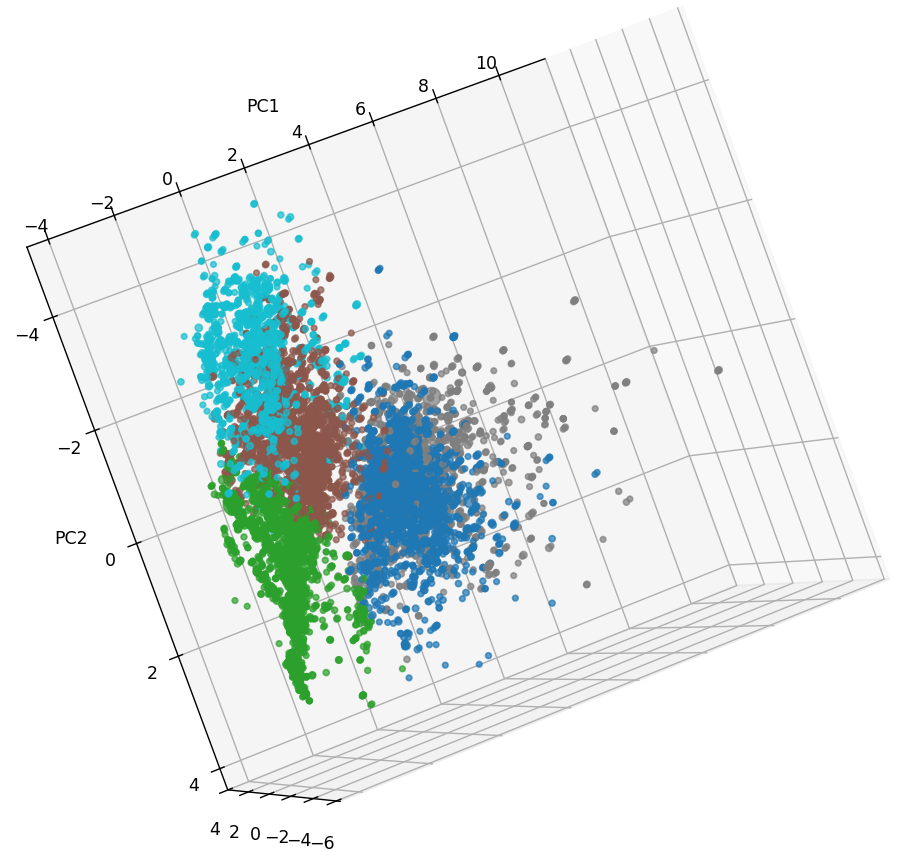


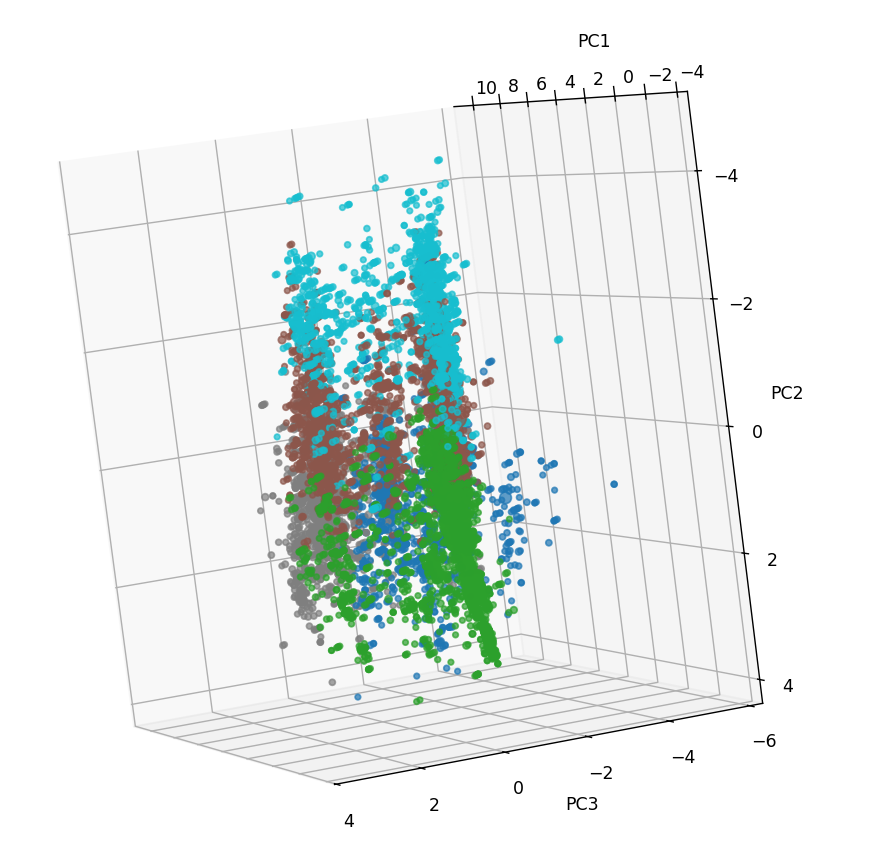
As we can notice in the graph, there are a number of samples that are relatively large and light blue – They represent samples with large tumor size and metastases. They are in close proximity, and located in the negative side relative to the PC2 component, and relatively centered around the other axis’s.

From Analyzing the entries of the PC2 vector, we noticed the largest values are matching to features “Surgery Sum” and “PR sensitivity”. A low sensitivity to PR doesn’t allow the treatment with progesterone hormone. Additionally, low “Surgery Sum” suggests the patient had low number of surgeries. Both factors might contribute to the increase of tumor size, and of metastases.

* We created the same PCA graph, while coloring dots according to a k=6 KMeans clustering:







As we can see in the graph, the data has a clear partitioning into different group in the 3d space which matches the resulting clustering.

The partitioning into clear separate groups might help to classify groups of patients with similar features, in order to deduce some properties of the tumor on all of the group, after performing some observation on some of it.

We believe some of the clustering derives from the way we performed preprocessing – many verbal features were converted into numerical values, creating discrete gaps along the matching axis’s. Therefore, we could expect the dots to be spaced in constant gaps along the matching axis, and also pass similar property to the lower dimension representation.