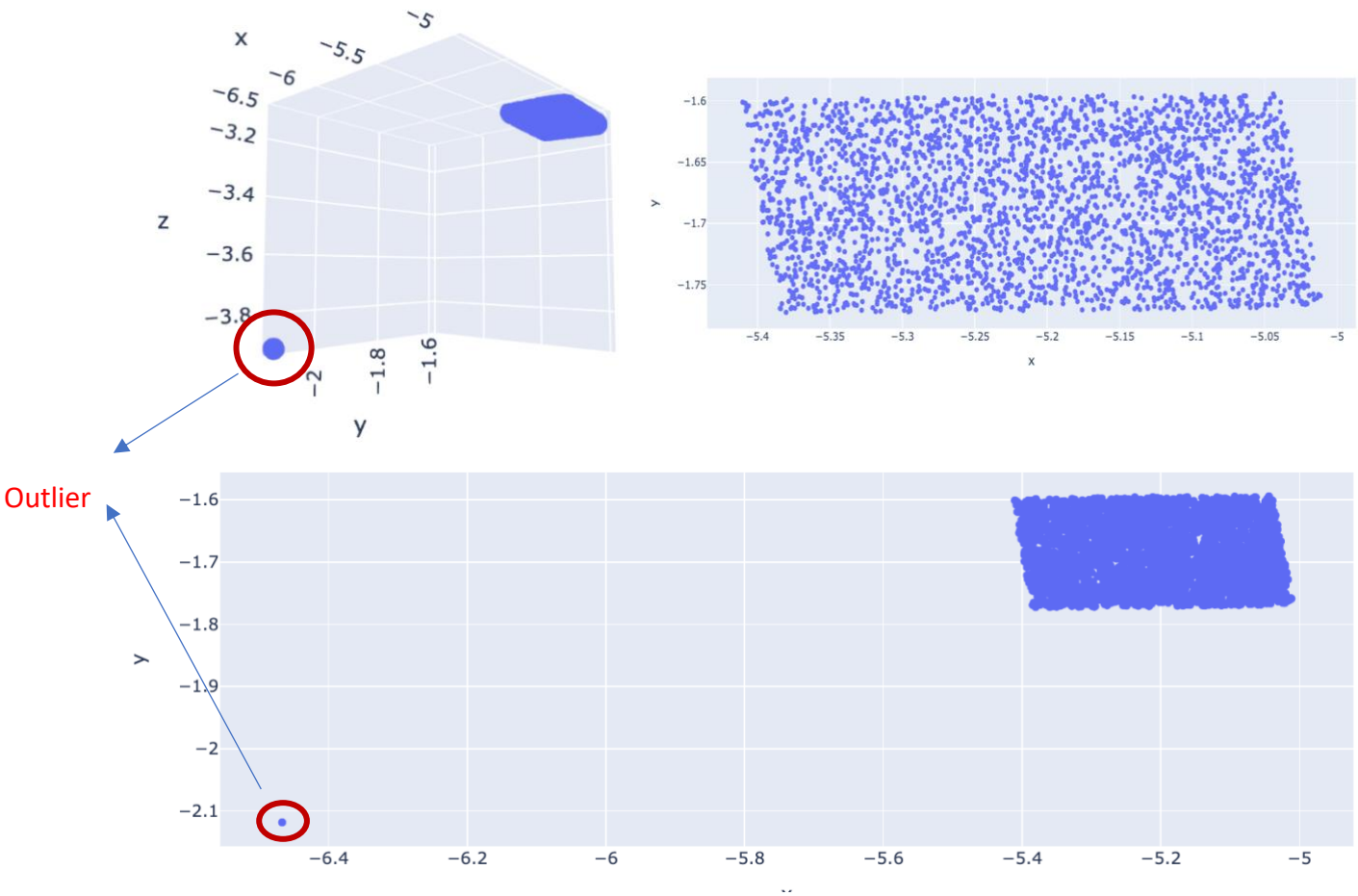
Final Report- Clustering Analysis for 6D Data using Spark

By: Srinidhi Manikantan

Cluster 1



The first shape I have identified is a **3D parallelogram**, **centered at [75, 75, 75, 75, 75, 75] position** (all rounded to the nearest integer). There is a total of **3001 data points** that belong to cluster 1, one of it being an **outlier** which I have identified it to be datapoint **[91, 92, 93, 94, 95, 96]**. We know that this parallelogram is a 3D figure based on the cumulative variance obtained for each PCA component:

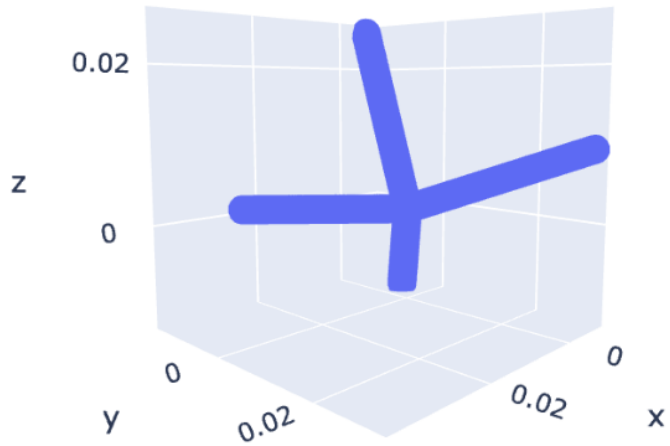
[0.81558969, 0.98493942, 1.0, 1.0, 1.0, 1.0]

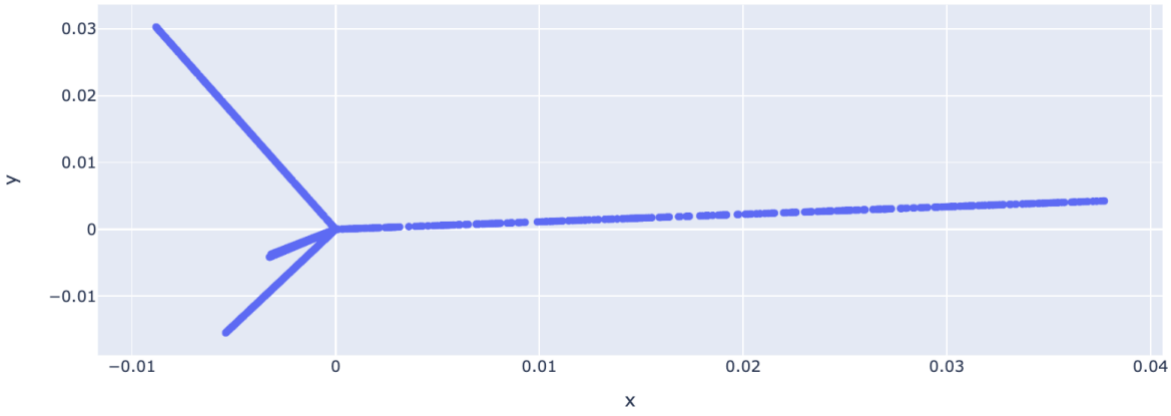
Given that the explained variance for the third and subsequent components is 1.0, we can deduce that the data could be represented in a 2D or 3D subspace of the original 6D space.

In 3D space, I have identified that the object has a length of 0.39 units, width of 0.18 units and height of 0.018 units. Looking at these values, we can see that the height is considerably smaller than the other two dimensions. Hence, we can consider it a **3D object with a quasi-2D characteristic due to the small height dimension**.

In terms of size, we know that **the length of the parallelogram is about x2 times the size of the width** (0.39/0.18 = ~2). My guess is that the actual size is about **~10units in length, ~5 units in width and about ~1 unit in height.**

Cluster 2





The second shape I have identified is a **6D "tripod" figure**, **centered at [0.08, 0.08, 0.08, 0.08, 0.08, 0.08] position** (all rounded to the nearest 2 decimal points). There is a total of **2400 data points** that belong to cluster 2. **No outliers** can be found for this cluster. We know that this tripod shape is a 6D figure based on the cumulative variance obtained for each PCA component:

[0.2456057, 0.44725014, 0.62943154, 0.79390144, 0.95498558, 1.0]

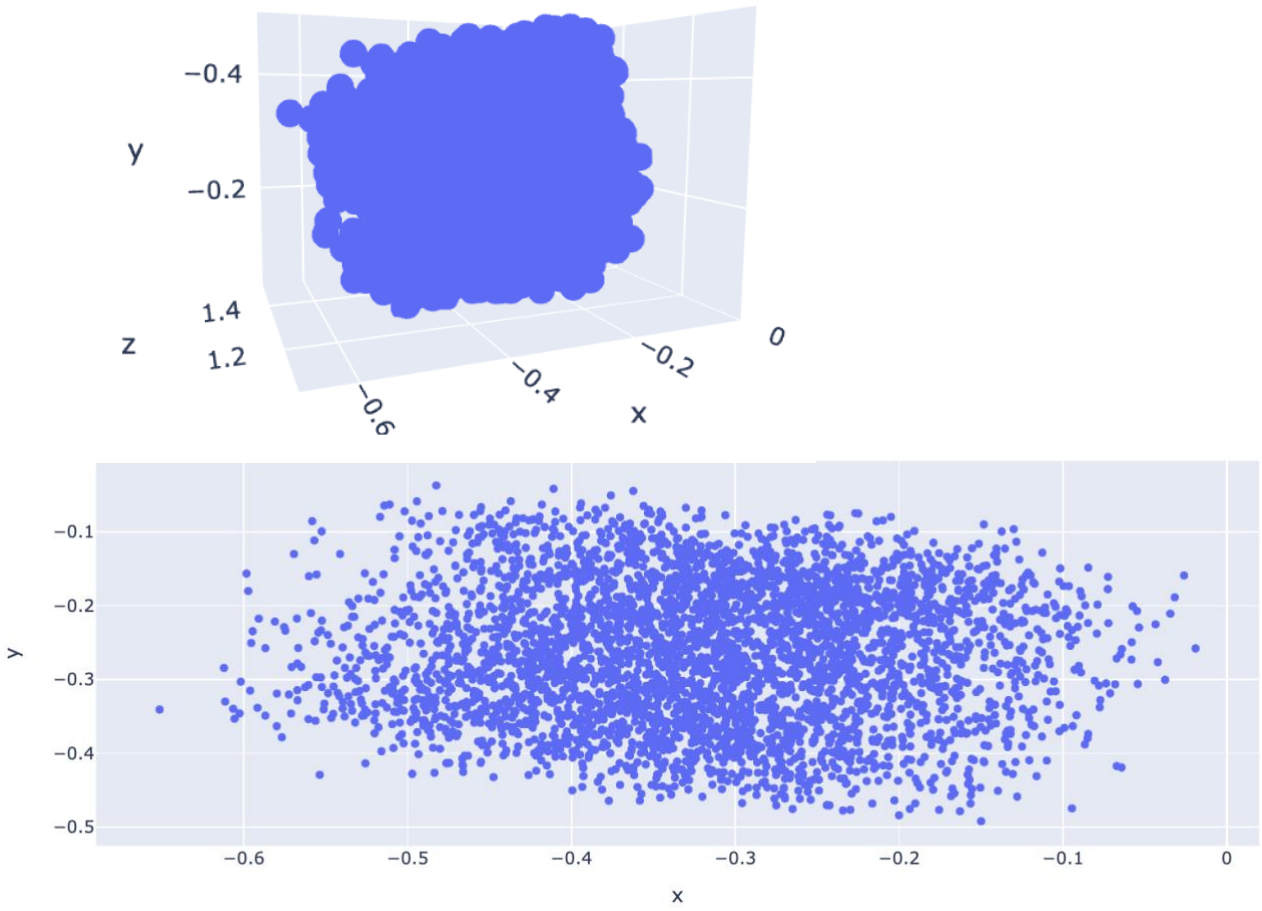
Based on the cumulative variance, we can conclude that the shape we have obtained for this cluster is six-dimensional. Each principal component contributes to a dimension, and all six principal components are needed to capture the entire variance in your data.

In terms of size – although the shape we have obtained above looks like the "legs of the tripod" have different sizes, I have identified that the **object equally spans across the 6**

different axes (i.e., the length/range of values it occupies along each axis is the same). I have also identified that this length is about **1 unit** for all 6 axes.

Given that the length/range of values the object occupies in each axis is roughly the same (~0.99units), this means that spread or variability of values is consistent across those axes.

Cluster 3



The third shape I have identified is a **6D hypercube/ cloud-looking figure**, **centered at [15, 80, 15, 80, 15, 80] position** (all rounded to the nearest integer). There is a total of **4000 data points** that belong to cluster 3. **No outliers** can be found for this cluster. We know that this "hypercube" shape is a 6D figure based on the cumulative variance obtained for each PCA component:

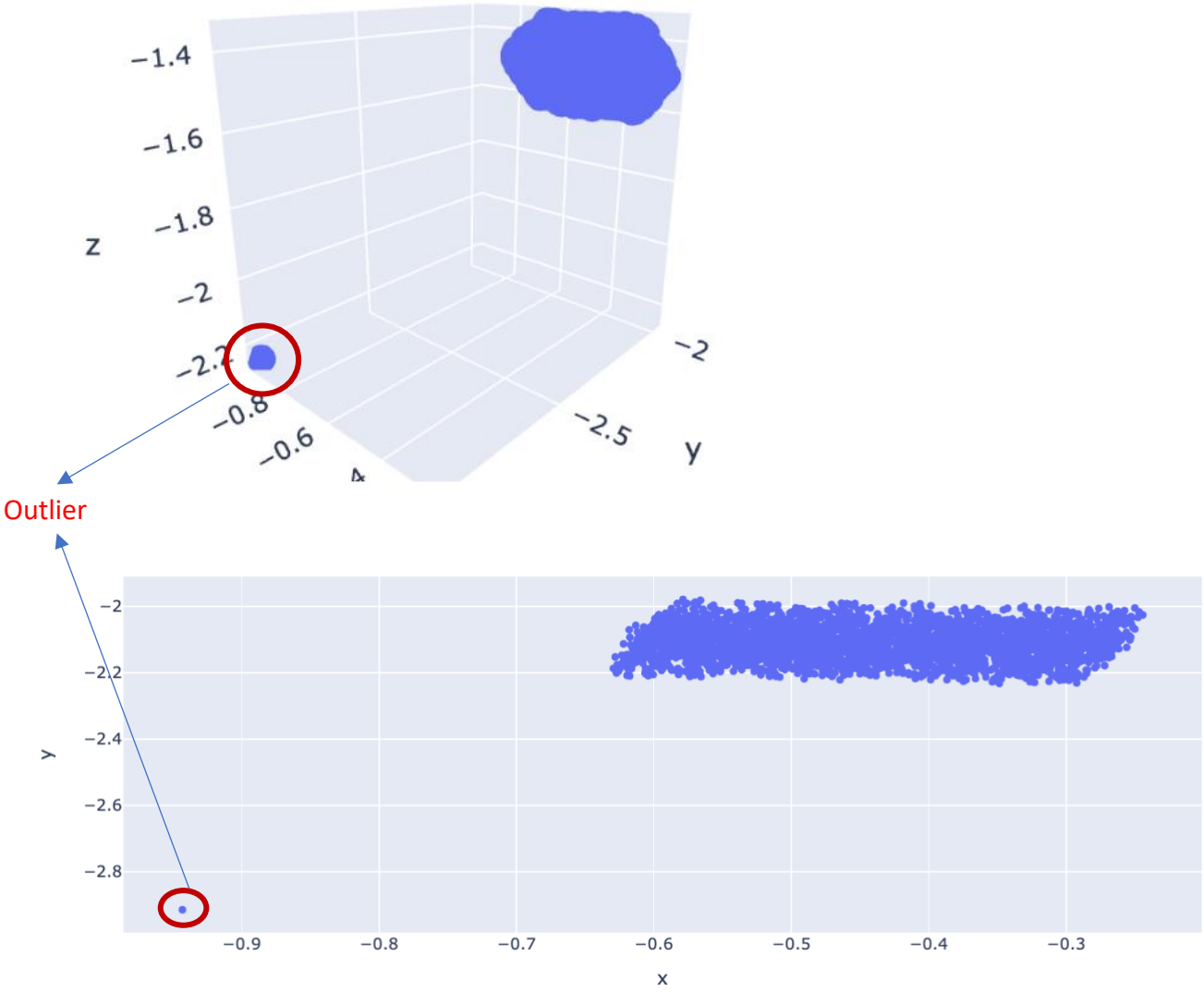
[0.23610388, 0.41357649, 0.5704253, 0.72211295, 0.86981003, 1.0]

Based on the cumulative variance, we can conclude that the shape we have obtained for this cluster is six-dimensional. Each principal component contributes to a dimension, and all six principal components are needed to capture the entire variance in your data.

In terms of size – although the shape we have obtained above looks like the length > width (due to different scales), I have identified that the **object quite equally spans across the 6**

different axes (i.e., the length/range of values it occupies along each axis is the same). I have also identified that this length is roughly around **15 units** for all 6 axes.

Cluster 4



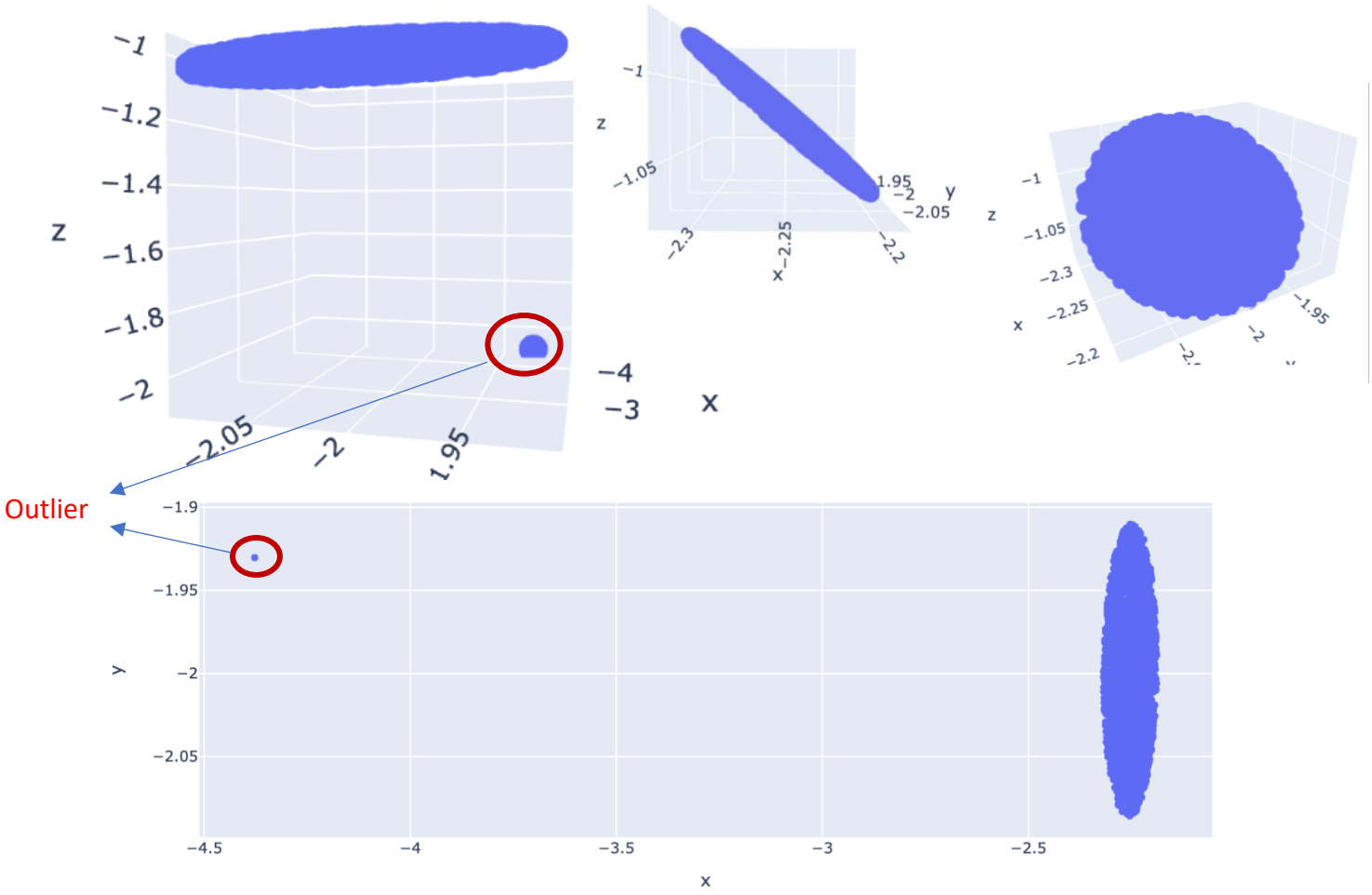
The fourth shape I have identified is a **3D parallelogram**, **centered at [25, 25, 25, 75, 75, 75] position** (all rounded to the nearest integer). There is a total of **3501 data points** that belong to cluster 4, one of it being an **outlier** which I have identified it to be datapoint **[11, 12, 13, 94, 95, 96]**. We know that this parallelogram is a 3D figure based on the cumulative variance obtained for each PCA component:

[0.65361961, 0.82866565, 0.99133456, 1.0, 1.0, 1.0]

Here, we can see that the explained variance for the third component is very close to 1.0 and is collectively able to explain 100% of variance in the data. This suggests that our object has a minimum of 3 dimensions. The fact that the cumulative variance values for the 4th, 5th, and 6th components are also 1.0 suggests that this object is effectively 3-dimensional, and introducing more dimensions does not contribute significantly to explaining the variance.

In terms of size, I have identified that the object in 3D space has a length of 0.38 units, width of 0.25 units and height of 0.22 units. From this, we can deduce that the **ratio of the length to the width of the parallelogram is about 3:2.**

Cluster 5



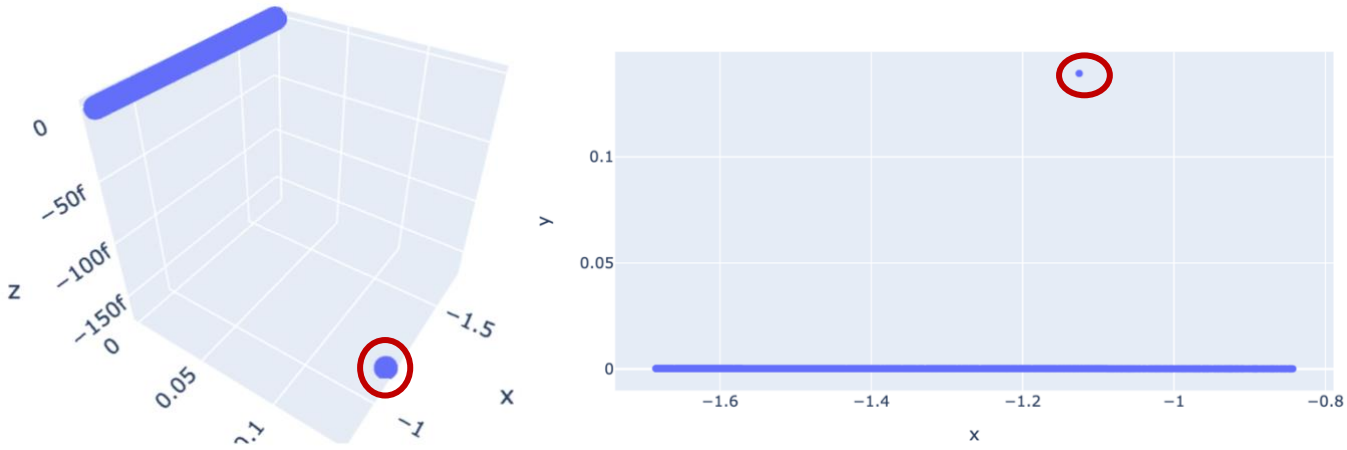
The fifth shape I have identified is a **3D disc**, **centered at [70, 60, 50, 40, 30, 20] position** (all rounded to the nearest integer). There is a total of **2501 data points** that belong to cluster 5, one of it being an **outlier** which I have identified it to be datapoint **[91, 92, 93, 14, 15, 16]**. We know that this disc is a 3D figure based on the cumulative variance obtained for each PCA component:

[0.47920818, 0.79350254, 1.0, 1.0, 1.0, 1.0]

Given that the explained variance for the third and subsequent components is 1.0, we can deduce that the data could be represented in a 3D subspace of the original 6D space.

In terms of size, the **diameter of the disc (which seems to span across the y axis), is about 3 units in 6D space. The orientation spans about 0.955 radian.**

Cluster 6



The sixth shape I have identified is a **1D line**, **centered at [15, 15, 15, 15, 15, 15] position** (all rounded to the nearest integer). There is a total of **2001 data points** that belong to cluster 6, one of it being an **outlier** (see the circled point above). However, when I use z-score method, there seems to be no detection of outliers. Even the furthest point identified from the mean does not identify this as an outlier. Hence, it is interesting to see that visually we can detect an outlier but not through statistical means.

We know that this line is a 1D figure based on the cumulative variance obtained for each PCA component:

[0.99983508, 1.0, 1.0, 1.0, 1.0, 1.0]

Here, we can see that the explained variance for the first component is very close to 1.0 and is collectively able to explain 100% of variance in the data. This suggests that our object has only one dimension. The fact that the cumulative variance values for the second and subsequent components are all 1.0 suggests that this object is effectively 1-dimensional, and introducing more dimensions does not contribute significantly to explaining the variance.

In terms of size, the **length of the line is about 10 units** in the 6D subspace. Given that the length/range of values the object occupies in each axis is roughly the same (~9.98 units), this means that **spread or variability of values is consistent across those axes.**