Data Mining - Exercise 1

Student Name	Student-ID
Marco Di Francesco	100632815
Loreto García Tejada	100643862
György Bence Józsa	100633270
József-Hunor Jánosi	100516724
Sara-Jane Bittner	100498554

a) The Distance values (min, max, average) logarithmically increase for all the norms, rapidly after the first increase of dimensions. The differences-measure behaves differently for various L-norms. For L0.5 and L1 the difference increases, for L2 it stays the same, and it decreases for L3. Here it can be seen that for a lower p, the difference still remains higher, compared to higher p values.Therefore, lower p-values should be chosen in case of higher dimensions. As the number of dimensions increases, the average difference between the extremes gets less and less significant. This is because for an extreme distance to be had, *d* features have to have values that are either all near or far from the features of **x**. As *d* increases, the probability of all features having this property gets smaller and smaller.

b) With increasing dimensions, the contrast value decreases, independently of the L-norm. This can be explained with the distance plots: here, with higher dimensionality the differences between the extremes become smaller. Therefore, the contrast value becomes smaller as well. While all Lp-norms degrade with increasing dimensionality, the degradation is much faster for the plots representing larger values of p.

c) As seen in the plots the choice of p does not have a noticeable impact at the relative contrast value, it only has the very little difference that with high p norms it tends to slightly quicker. However, this difference is very small and generally the contrast value drops quickly and rapidly with increasing dimensions and after this drop continues further to decrease slightly with increasing dimensions. This observation is independent from the chosen p. This shows that the l-norms suffer from the course of dimensionality similarly, because the distances appear more similar with increasing dimensionality.

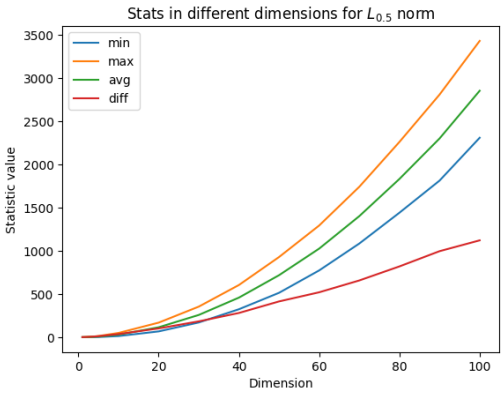
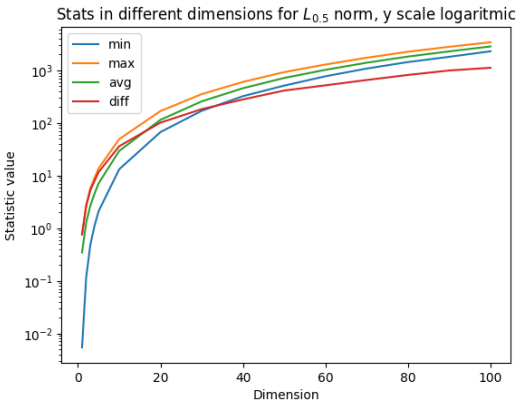
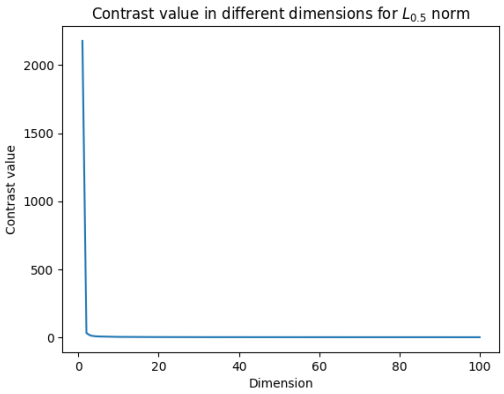
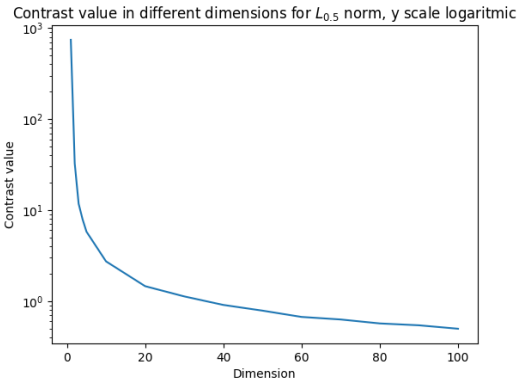
d) The *curse of dimensionality* refers to the tendency of high dimensional data behaving differently than low dimensional data. In our case, this curse means that distance measures become less informative as the number of dimensions increase. This is a problem for clustering high dimensional data, it requires a measure of distance or similarity.

Overview of Plots:

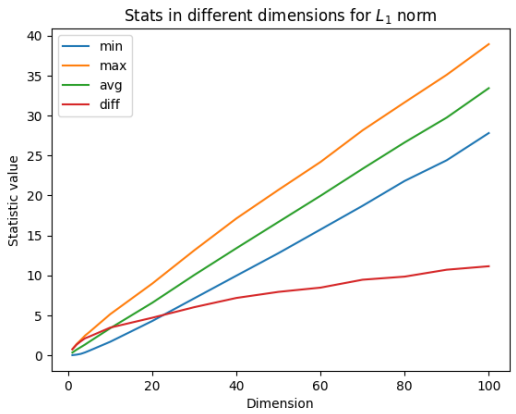
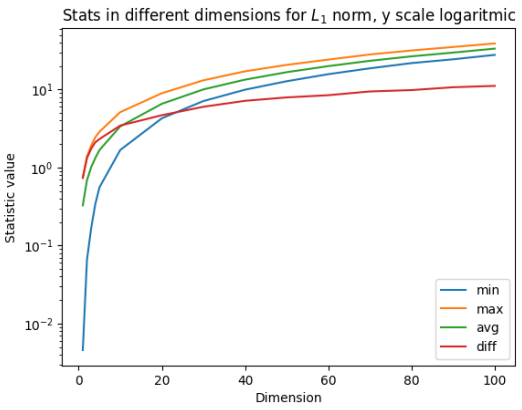
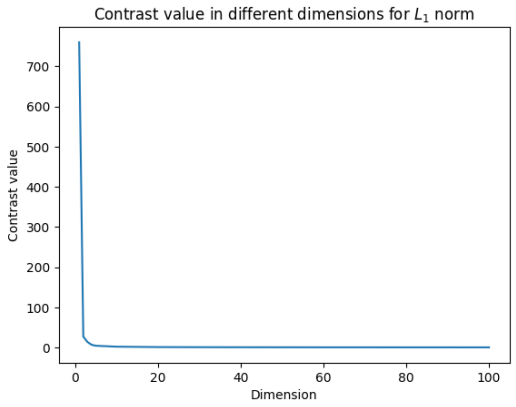
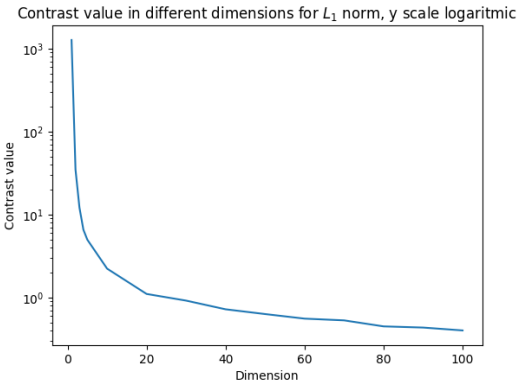
Avg = Average Values (Minimum, Maximum, Mean, Difference)

CV = Contrast Value

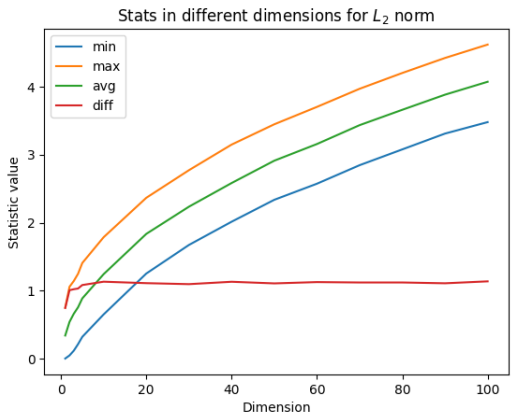
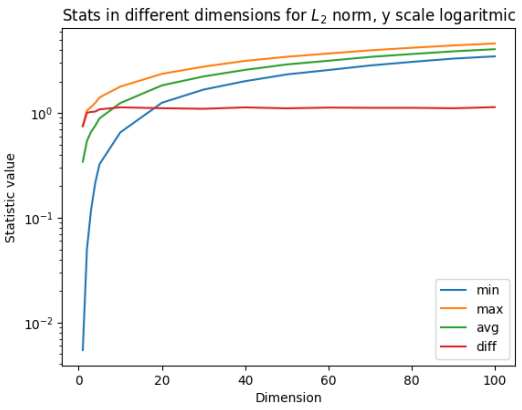
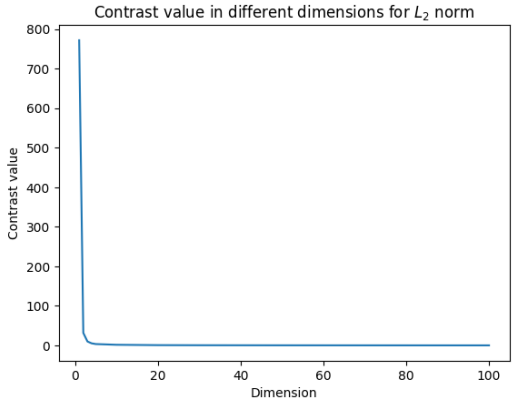
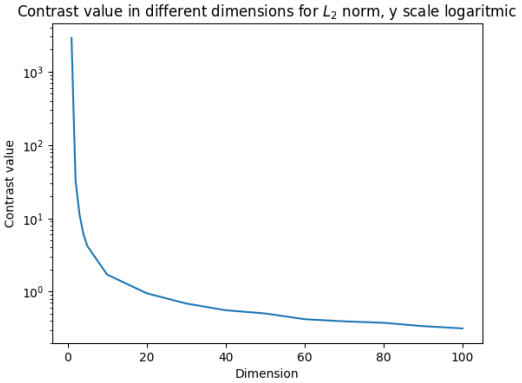
1. **p = 0.5**

	Linear	**Log**
Avg		
CV		

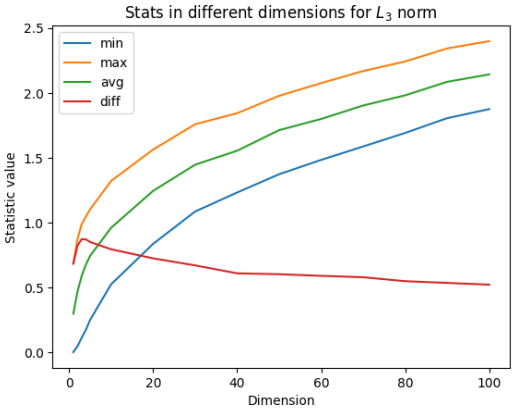
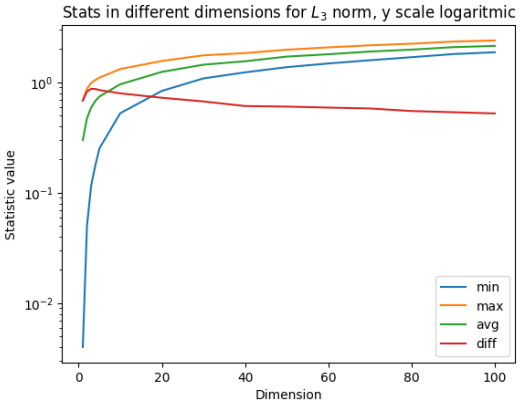
2. p = 1

	Linear	Log
Avg		
CV		

3. p = 2

	Linear	Log
Avg		
CV		

4. p = 3

	Linear	Log
Avg		
CV