Density-Based Clustering
Exercises

Exercise 1. (and Distances in Geo-Spatial Data)
Locations on a planet (e.g. earth) are expressed in latitude and longitude.

1. Which distance function should be applied to measure the distance between two locations?
2. Find a suitable candidate in Python!
3. How do you have to transform the geo-spatial coordinates in order to apply the distance function?

Exercise 2. (or Clustering Geo-Spatial Data)
Use density-based clustering and k-means on a dataset of geo-spatial data.

1. Load the file `yellow_tripdata_2016-03.csv`, which is available under a public domain license on Kaggle.
2. In this task we focus only on the pickup locations of taxi trips.
3. To avoid long runtimes, use only a fraction of the dataset. This can be created using the method `sample(frac, random_state)` on a data frame. For this task, use only 1 per mill of the available data and use a random seed of 1.
4. Plot the data using a regular scatter plot.
5. Use DBSCAN to cluster the data with these parameters: $\varepsilon = 0.000005$, MinPts $= 20$. (Keep your considerations of the previous task in mind!) Plot the resulting clusters.
6. For comparison, run k-means on the same data, using the number of clusters you got from DBSCAN as k and plot the resulting cluster.
7. Use the folium based method `create_map` from `07_e1_folium` to visualize you clustering on a real map of the area.
8. Discuss the differences of the two clusterings. Take particular care of the use case of a taxi service that plans the allocation of its fleet.

Exercise 3. (Optics)
Use the setup of the previous task. Create a clustering using OPTICS with MinPts $= 20$, $\xi = 0.005$, $\varepsilon_{max} = 0.00005$, and a minimum cluster size of 0.5% of the data.