

## Assignment: Drone delivery system

### Learning goals

In this assignment, you:

1. learn to conduct cluster analysis.
2. improve your data manipulation skills in Python.



Image: HadasBandel, via Wikimedia Commons, under [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

### Assignment

In near future, drones (unmanned aerial vehicles, see the image above) may carry out routine delivery tasks. To get an idea of the concept, have a look at Amazon Prime Air web pages: <https://www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011>

Your goal is to find ideal locations for a set of drone depots, based on the coordinates of the clients.

The simulated data set is in the **Methods/Data/Drone delivery** folder at the **Documents** tab of the course's Oma workspace. Each row starts with a customer identifier, followed by the geographical locations of nearly 6,000 clients.

1. Visualize the client locations by making a two-dimensional scatterplot.

Hint: you can use Seaborn <https://seaborn.pydata.org/generated/seaborn.scatterplot.html> or Matplotlib: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/visualization.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html). Can you give a geographic interpretation for what you see?

Using k-means clustering, find optimal locations (i.e. x and y coordinates) for three drone depots. Each depot should serve its surrounding clients.

Hint: The centroids serve as the depot locations. You will later need to change the number of depots, so design your program in such a way that you just need to modify a single value to do that.

2. Attach the information on the closest depot to each client. That is, generate a data frame that is similar to the original one with the exception that it has an additional column that contains the identifier of the depot nearest to the client. Print the first 10 rows of the new data frame.
3. Make a scatterplot that uses three different colours. The markers with the same colour are served by the same depot.

Hint: Re-check the web page(s) mentioned in the first task.

4. Play with the number of depots. What are the optimal locations for 10 depots, for example? Do you see a difference in the computation time when the number of depots increases?
5. Replace k-means with agglomerative hierarchical clustering and explore it with various depot numbers. What are your observations?

## **Deliverables**

Your deliverable should include both the Python codes and the results as well as relevant discussion.

Please submit the answer as a downloaded HTML document. In Jupyter workbook, select **File / Download as / HTML (.html)**.