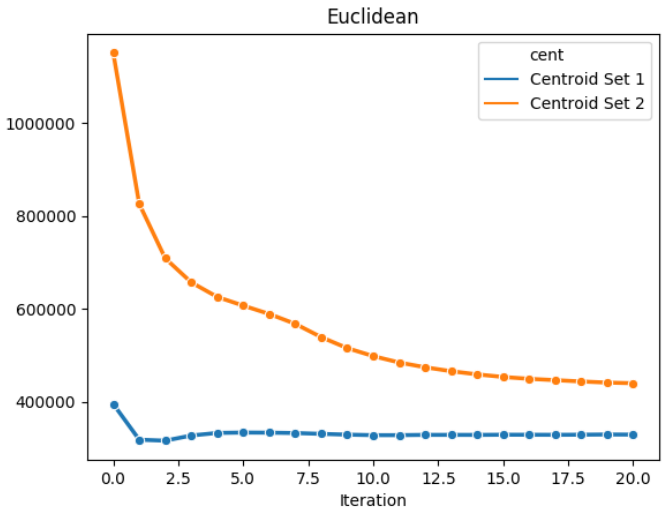
BIG DATA MINING, HW2, RESULTS

RORY FLYNN

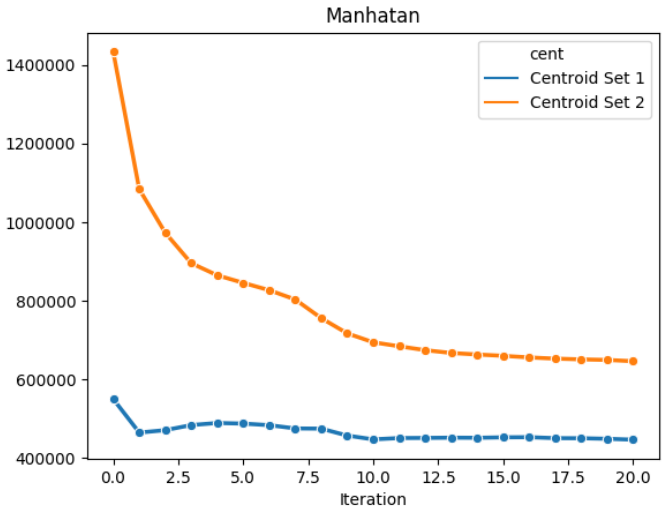
Part 1

Deliverable 1.1: The code as a Jupytier Notebook (.ipynb) or a .txt file (30 points). The code is in one txt file which will be submitted along with this document. It is used with a driver python file also included as a txt.

Deliverable 1.2: A plot of cost vs. iteration for two initialization strategies for a.1 (5 points). Below is a plot of the Euclidean Distance Cost vs iterations, for random cluster initialization(c1), and max-distance cluster initialization(c2).



Deliverable 1.3: A plot of cost vs. iteration for two initialization strategies for b.1 (5 points). Below is a plot of the Manhattan Distance Cost vs iterations for random cluster initialization(c1), and max-distance cluster initialization(c2).



Deliverable 1.4: Your answer for questions a.2 and b.2 (10 points.

- **a.2:** If you want to save time and computation power while optimising you clustering in terms of the cost metric; then our results indicate that you should chose random cluster initialization(c1), over max-distance cluster initialization(c2). One need only look at the graphs above to see a clear example of the c1 method. The max-distance centroids not only start with higher cost, they seem to reach convergence well above the random centroids.
- **b.2:** Whether the cost is calculated in Euclidean or Manhattan distance the answer remains the same. Random is clearly superior to max distance in this case. It is notable that both metrics produce similar results, with only trivial differences. This may be more interesting if cosine distance was used instead of Manhattan.