מטלה 4:

http://scikit-learn.org/stable/modules/clustering.html

השוואות של:

**שיטות אתחול K mesans:**

**Forgy method** -  randomly chooses *k* observations from the data set and uses these as the initial means

**The Random Partition** method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points.

**K-means++:**

1. Choose one center uniformly at random from among the data points.
2. For each data point *x*, compute D(*x*), the distance between *x* and the nearest center that has already been chosen.
3. Choose one new data point at random as a new center, using a weighted probability distribution where a point *x* is chosen with probability proportional to D(*x*)2.
4. Repeat Steps 2 and 3 until *k* centers have been chosen.
5. Now that the initial centers have been chosen, proceed using standard [*k*-means clustering](https://en.wikipedia.org/wiki/K-means_clustering).

השוואות שיטות עדכון:

Mini Batch K-Means

K-means

Elkan

מדדים:

1. לקחת מידע מתויג ולראות כמה זה קרוב לתיוג האמיתי – מדדים למידע מתויג
2. מדדי עבודה 2 ל-clustering
3. זמן ריצה

דרכי הערכה:

1. עד להתכנסות (יש כמה פרמטרים להתכנסות)
2. או להגביל מספר צעדים
3. התנהגות עבור – k שונים

**מדדים סטטיסטיים**

**ניתוח רגישות**

**אילוסטרציה**

**לכל מדד – לעשות ממוצע של המדדים**

**להשוות למימוש של SK-Learn**