**מדעי הנתונים ובינה עסקית תרגיל מס' 4**

בעבודה זאת ממשנו תהליך המבצע clustering לקובץ נתונים. התהליך מורכב מהכנת הנתונים, חלוקת הנתונים לאשכולות והצגת תוצאות.

הקוד מחולק למחלקות בהתאם לתהליכים:

1. **Preprocess** – המחלקה מבצעת הכנה של הנתונים בעזרת שיטות הבאות:

* **\_\_init\_\_(self, path)** בנאי של המחלקה, מקבל נתיב לקובץ xlsx ויוצר ממינו שדה מסוג data frame המכיל נתונים של הקובץ.
* **clean\_na(self)** ממלאה נתוני תכונות חסרים בממוצע של נתוני התכונה.
* **standardization(self)** מבצעת נורמליזציה של הנתונים לערך הסטנדרטי עבור כל תכונה.
* **aggregate\_by\_country(self)** עבור כל מדינה מבצעת חישוב ממוצע של כל תכונה לפי שנים ויוצרים data frame חדש מקובץ לפי מדינות וללא תכונה ‘year’.

1. **Clustering** – המחלקה מבצעת חישוב של האלגוריתם K-Mean Clustering לפי פרמטרים של משתמש והנתונים של התרגיל בעזרת שיטות הבאות:

* **\_init\_\_(self, data\_frame)** בנאי של המחלקה, מקבל לשדה data frame שאחרי שלבי preprocess.
* **activate\_k\_means\_algorithm(self, n\_clusters, n\_init)** מקבלת כפרמטרים כמות האשכולות וכמות הפעמים שהאלגוריתם ירוץ על centroid seeds שונים. השיטה מבצעת חישוב לפי האלגוריתם ושומרת את התוצאות בשדה data frame.
* **create\_scatter\_generosity\_social\_support(self)** מציירת scatter לפי תוצאות של האלגוריתם עבור ה-data frame.
* **create\_country\_map(self)** מציירת Choropleth Map לפי תוצאות של האלגוריתם עבור ה-data frame.