# Homework 1 – Business Intelligence

## חלק 1 – אוסף נתונים

בחלק זה, עליכם למצוא אוסף נתונים אשר עליו תבצעו את כלל המטלות לאורך הסמסטר.

אוסף הנתונים צריך לקיים את הדרישות הבאות:

1. יש להשתמש בנתונים במודל הטבלאי (excel, SQL,...)
2. להכיל (לפחות) 2 טבלאות שונות אשר קשורות אחת לשנייה. במידה ומצאתם רק dataset אחד שמעניין אותכם, **תמיד ניתן לנרמל את המסד-נתונים ולהפוך אותו למספר טבלאות**.
3. להכיל תכונות מסוגים שונים. למשל, מספרים שלמים, מספרים שברים, טקסט, מייל, וכו׳.
4. יש להשתמש בraw-data בלבד, ללא שום ניקוי מקדים.

מקורות מידע

- רשימה של אוסף מידע עבור ניתוח עסקי:

https://sqlbelle.wordpress.com/2015/01/16/data-sets-for-bianalyticsvisualization-projects/

מומלצים מתוך האתר הנ״ל:

Sports, Movie, Music, Stocks

- אתר Kaggle המציע אוסף מידע למטרות שונות:

https://www.kaggle.com/datasets

- אתר UCI (מצויין ומכיל המון אוסף נתונים):

https://archive.ics.uci.edu/ml/datasets.php

מומלץ: להשתמש במסנני האתר (תפריט בצד שמאל) ולמצוא את אוסף הנתונים אשר מעניין אותכם.

- אוסף נתונים עבור covid-19:

https://www.ecdc.europa.eu/en/covid-19/data

- **מומלץ!** במידה והנכם עובדים בתחום, להשתמש באוסף נתונים מהעבודה על מנת להתמודד עם real-life-situations

## חלק 2 – שאלות מחקר / שאלות עסקיות

1. נסחו 2 שאלות מחקר / שאלות עסקיות עבור אוסף הנתונים שמצאתם. שאלת מחקר אחת צריכה להיות supervised והשנייה unsupervised.
2. עבור כל שאלת מחקר, הגדירו את הKPIs והסבירו מדוע היא SMART (יש לפרט במשפט עבור כל סעיף).

## חלק 3 – הבנת הנתונים

השתמשו בכלי לניתוח סטטיסטי פשוט (למשל אקסל, R, python) והציגו תחקור נתונים.

תחקור נתונים הוא "התמונה הראשונה" שאתם רואים מאוסף הנתונים וצריך להכיל:

1. מדדי פיזור של התכונות (ממוצע, סטיית תקן, רבעונים)
2. תלויות וקשרים (קורלציה)
3. כל מדד סטטיסטי נוסף אשר יכול לעזור לכם לקבל הבנה טובה יותר על הנתונים.
4. עבור שאלת הsupervised, מצאו את האנטרופיה של כל אחת מהתכונות.
5. עבור 2 התכונות בעלות האנטרופיה הנמוכה ביותר, חשבו את מדדי הgini-index והinformation-gain.
6. מה ניתן להסיק מתחקור הנתונים? רשמו ב2 משפטים קצרים.