Linear Regression

בתרגיל זה נעבור עם מאגר הנתונים World Happiness report 2020

<https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2020.csv>

נמצא גם בחומרי הקורס

להלן הסבר עבור כל העמודות בקובץ **2020.csv** מתוך מאגר הנתונים *World Happiness Report*:

1. **Country name** – שם המדינה שנבדקה בדוח.  
   *דוגמה:* Israel, United States, Germany.
2. **Regional indicator** – האזור הגיאוגרפי שאליו המדינה משתייכת, כמו מערב אירופה, דרום אסיה, אפריקה שמדרום לסהרה וכו'.  
   *דוגמה:* Western Europe, Sub-Saharan Africa.
3. **Ladder score (Happiness Score)** – ציון האושר במדינה, שהוא ערך ממוצע שמבוסס על סקרים שבהם אנשים דירגו את רמת האושר שלהם בסולם מ-0 עד 10.  
   *דוגמה:* 7.8, 5.4.
4. **Logged GDP per capita** – התוצר המקומי הגולמי (GDP) לנפש במדינה, לאחר הפעלת לוגריתם, כדי להתאים את קנה המידה של הנתונים.  
   *דוגמה:* 10.5, 9.2.
5. **Social support** – מדד לתמיכה חברתית, המבוסס על השאלה: "האם יש לך מישהו לפנות אליו בעת צרה?" ונמדד בערכים מספריים.  
   *דוגמה:* 0.95, 0.80.
6. **Healthy life expectancy** – תוחלת החיים הבריאה בלידה, כלומר מספר השנים הצפויות שאדם יחיה בבריאות טובה.  
   *דוגמה:* 67.5, 55.3.
7. **Freedom to make life choices** – מדד לחופש בחירה אישי, שמייצג את תחושת האזרח לגבי חופש הבחירה בחייו.  
   *דוגמה:* 0.60, 0.45.
8. **Generosity** – מדד הנדיבות, שמבוסס על שאלות לגבי תרומה כספית לאחרים בשנה האחרונה.  
   *דוגמה:* 0.20, 0.15.
9. **Perceptions of corruption** – מדד לתפיסת השחיתות במדינה, המתבסס על תחושות הציבור לגבי מידת השחיתות בגופים ממשלתיים ועסקיים.  
   *דוגמה:* 0.10, 0.35.

**שאלות**

**מבוא:**

1. מחק פיזית את העמודות ב DF שלא מופיעות בתיעוד מעלה.
2. כמה שורות וכמה עמודות יש כעת במאגר ?
3. האם יש במאגר ערכי Nan, אם כן, טפל בהם
4. בדוק עבור ציון האושר ערכי מינימום, מקסימום וממוצע. האם נראה הגיוני?

**ויזואליזציה: דאג לכך שהגרפים יהיו אפשריים לקריאה**

1. הצג את מדד האושר לפי יבשת. לאיזו יבשת יש את מדד האושר המקסימאלי? למי המינימאלי?
2. הצג את מדד האושר לכל מדינה ממוין לפי ערך מדד האושר בסדר יורד. באיזה מקום נמצאת ישראל?
3. הצג את מדד האושר מול GDP לנפש עם משתנה hue עם ערך יבשת.
4. המחש את pairplot. לאיזה משתנה יש יחס ליניארי יורד עם מדד האושר.

**קורלציות:**

1. לאיזה שלושה משתנים יש את הקורלציה הגבוהה ביותר עם מדד האושר?
2. מיהי המדינה עם מדד האושר הגבוהה ביותר? האם תוכל להסביר מדוע בעזרת המשתנים שמצאת בסעיף קודם?
3. הראה את הקורלציות שמצאת באמצעות מפת חום

**רגרסיה ליניארית:**

1. אמן את מודל הרגרסיה
2. מהם ערכי המקדמים? מה ערך הקבוע?
3. הערך את איכות הרגרסיה הליניארית באמצעות R בריבוע ובאמצעות RMSE