להלן התרגיל:

התרגיל יבוצע בשפת פייתון.

אנחנו ממליצים על התקנת Anaconda

ופיתוח המטלה ב – Jupyter (מחברת)

לפני שאתה מתחיל – תקרא קצת על פייתון

ועל החבילה שיש לעשות בה שימוש Pandas.

מדובר אובייקט המרכזי אנחנו עובדים על מנת לנתח מידע, לעבד ולהריץ במודלים.

האתר שבו יש את הדרך ללמוד הכי טוב נמצא ב - <https://pandas.pydata.org/>

מאמר לא רע אתה יכול למצוא ב - <https://medium.com/@oren_94714/what-you-must-know-about-pandas-hands-on-scenario-154ed7391700>

מטרת התרגיל:

טיפול במידע והבנה של אובייקט מרכזי בפייתון

חשוב שתקרא קצת ותבין את תהליך ההתקנה של החבילה. תהליך ההתקנה הוא מאוד פשוט והוא זהה לכל חבילה:

pip install pandas !

אז אחרי שקראת והתקנת להלן התרגיל:

1. לפניך שני קבצים: אחד csv   והשני xlsx .
2. עלייך לטעון את שני הקבצים לשני DataFrame שונים.
3. עלייך להציג (להדפיס) לפני הטיפול במידע את המידע הבא:
   * כמות עמודות ושורות בכל קובץ
   * את שתי השורות הראשונות
   * את השורות האחרונות
   * חמש שורות sample רנדומלי.
   * לכל אחד מהקבצים יש ״בעיות״ אחרות. חלק מערכים בעמודות ריקים.
   * בחלקן אתה צריך ליצור עמודה.
   * לאחר מכן עליך לאחד בין שני הקבצים ולהציג DataFrame שמאחד בין הנתונים.

משימת בונוס:

* + חלק אחרון עליך לבנות מודל PCA שנותן ציון אחד לכל שורה ושעושה שימוש בכלל הנתונים שאיחדת.
  + יש לקרוא על מודל PCA – הפחתת מימדיות - <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>

המלצות:

* + טפל בכל קובץ בנפרד.
  + תוודא שכשאתה מאחד את הקבצים אתה שומר על כמות העמודות שאתה מצפה לה.

כל שאלה אנחנו זמינים 😎

בהצלחה