מבוא לניתוח נתונים – תרגיל מספר 3

- תאריך הגשה: 4.5.21
- ההגשה לבד או בזוגות. במקרה של הגשה בזוגות -רק אחד מבני הזוג מגיש, אך יש להקפיד לרשום את השם ות.ז. של שני הסטודנטים שהגישו. מי שלא רשום על אף עבודה לא יקבל ציון, ולא יוכל להצטרף לעבודה בדיעבד.
- במידה וסעיף מסוים לא רץ – הציון על סעיף זה הינו אפס.
- יש להראות את הקוד דרכו הגעתם לפתרון. סעיף עם תשובה נכונה אך ללא קוד יקבל ציון אפס.
- התרגיל הינו אישי וניתן לבצעו בזוגות בלבד. אין להעביר פתרונות או חלקי פתרונות בין סטודנטים בכיתה. העברת התרגיל על כל המשתמע מכך לאנשים אחרים שאינם רשומים לקורס אסורה. כל הנ"ל הינם עבירות משמעת אשר יועברו לבחינת ועדת המשמעת של האוניברסיטה.
- הקפידו להגיש עבודה ברורה עם הסברים. בכל שלב, יש להסביר בבירור על איזה סעיף אתם עונים. על סעיף ללא הסברים יורדו לפחות חצי מהנקודות.
- יש להגיש קובץ אחד בלבד, מסוג ipynb.
- איחורים בהגשה מכל סיבה שאינה נמצאת בתקנון הרשמי של האוניברסיטה: על כל יום איחור בהגשה יורדות 5 נקודות בציון התרגיל באופן אוטומטי.

הניחו כי אתם מנתחי נתונים עבור חברת יו-טיוב העולמית. החברה העמידה בפניכם חלק מנתוני הצפייה בשלוש מדינות – ארה"ב, הודו ויפן. קיראו את הדאטה מכאן:

https://github.com/nlihin/data-analytics/tree/main/datasets/youtube

1. (10%) מהו הסרטון עם מספר הצפיות הגבוה ביותר? מתי הוא פורסם ובכמה צפיות זכה? עליכם להציג אך ורק את הנתונים הללו: title, views, publish_time
2. (10%) הציגו את שלושת הערוצים עם מספר הצפיות הגבוה ביותר. עליכם להציג אך ורק את הנתונים הללו: channel _title, views (שלוש שורות שלהם).
3. (20%) מהי הקורלציה ע"פ spearman בין מספר ה likes, dislikes, views, comment_count?
4. (10%) צרו heatmap עבור סעיף 3.
5. (20%) צרו scatterplot להראות את שתי הקורלציות הכי חזקות שמצאתם בסעיף 3 (במידה ויש תיקו, ביחרו שתיים מתוך ההכי חזקות).
6. (20%) הציגו את כמות הצפיות (views) בכל קטגוריה (category_id) עבור השנים 2017-2018. אין להציג מידע נוסף (למשל – לא להציג גם את הצפיות ב2016, או את כמות ההערות). רמז: תחילה עליכם לחלץ את השנה ולשמור אותה בעמודה נפרדת. אחר כך מומלץ להשתמש ב pivot table.
7. (10%) נקודות מתנה. יורדו במידה ויוגש ערעור על הציון.

בהצלחה!!!

צוות הקורס