

ניהול מידע מבוצר - פרויקט חלק א'

חלק 1

שאלה 1.1

בשביל לסנן את העמודות שעומדות בתנאים כדרושים, בחרנו להוסיף את העמודות הבאות:

- בטבלה Demographic data נוסיף את עמודת gap_age - הפרש גילאים בין המבוגר הראשון למבוגר השני. השתמש בעמודה זו בשבייל התנאי השלישי, שכן נדרש לבדוק בו האם הפרש הגילאים הוא גדול/שווה ל-9, וכן נוכל פשוט לבצע filter על העמודה החדשה.
- בטבלה Daily program data נוסיף את עמודת weekday – היום בשבוע (בפורמט string קצר) שבו התוכנית עלתה לאויר. השתמש בעמודה זו בשבייל התנאי הרביעי, שכן נדרש לסנן רק את הרשומות שהתקיימו ביום שישי.
- בטבלה Daily program data נוסיף את עמודת monthday – היום מתוך התאריך (-1-31) שבו התוכנית עלתה לאויר. השתמש בעמודה זו בשבייל התנאי הרביעי, שכן נדרש לסנן רק את הרשומות שהתקיימו ביום ה-13.
- בטבלה Demographic data נוסיף את עמודת numeric_income – נמיר את האותיות D-A שהופיעו בעמודה זו למספרים נומריים כפי שהגorder בתרגום למספרים 13-10. נדרש את כל ההכנסות בתור מספרים על מנת לחשב ממוצע הכנסה, שכן בתנאי החמישי נדרש לסנן רק את המשפחות להן הכנסה מתחת לממוצע.
- בטבלה Daily program data נוסיף את עמודת genre_array – המטרה היא להפריד בין המחרוזות שהופרדה ב",", לערך של ז'אנרים, כדי שנוכל לבדוק בצורה נוחה את תנאי 6.
 - בעת עיבוד הנתונים והוספת הטבלה רצינו לחלק את הרשומות "הרלוונטיות" כלומר, רשומות אשר לפחות אחד מהז'אנרים שלחן שייר לרשימה הנתונה. בעת ביצוע עיבוד זה, נעזרנו בצ'אט GPT.

שאלה a 1.2

נציג CUT יתרוןות וחסרונות של השיטה שהוצעה בשאלה (כלומר איחוד 4 הטעלאות לטבלה אחת):

יתרוןות :

- פשוטות – מנהלים Data Frame ייחד המאפשר פשוטות בעקבודה עם תנאים מורכבים המשלבים שדות מספר טבלאות שונות.
- בדיקת התנאים על הרשומות היא נוחה יותר, שכן לכל רשותה ניתן לבדוק "בכט אחת", את כל 7 התנאים.

חסרונות :

- בזוז זיכרון ומשאבים – הטבלה המאוחדת תהיה מאוד גדולה ותכיל הרבה מידע, יתרן שהרבה מהמידע יהיה מיותר.
- כפיליות מיותרות – לאחר איחוד כל 4 הטעלאות, ישמר הרבה מידע כפול (לדוגמא עברור כל שידור, ישמרו רשומות כמספר מופעי הצפייה, והמידע על השידור יופיע המeon פעמים).

