**עבודת אמצע- רגרסיה**  
**1.** **נכון.** מתאם חלקי מיועד למצב שבו רוצים לבדוק את הקשר ה"נקי" שבין שני משתנים בניכוי משתנה/ים אחר/ים, והוא מבטא את הקשר הייחודי בין שני משתנים, שלא קשור למשתנה שלישי (או כמה) שנוכה מכל אחד מהם. כלומר- מתאם חלקי הוא בעצם מתאם פירסון בין 2 ציוני הפרש שנוכו מהם אותם המשתנים. בעצם, אם רוצים לחשב את המתאם שבין x ל-y בניכוי z, ניתן לחשב את המתאם שבין חלקו של x שלא מוסבר על ידי z, לבין חלקו של y אשר לא מוסבר על ידי z. בונים נוסחה לניבוי x על ידי z. מחשבים לכל נבדק את החלק ש-z לא הצליח להסביר על ידי חיסור החלק ש-z הצליח להסביר (הx המנובא של כל נבדק), מציון x האמיתי של כל נבדק. בונים נוסחה לניבוי y על ידי z, ועושים את אותה הפעולה על מנת למצוא את חלקו של y אשר z לא "הצליח" לנבא. לבסוף, מחשבים מתאם פירסון בין שני ציוני ההפרש הללו, אשר משקף את הקשר שבין x לבין y בניכוי z.

**2. לא נכון.** נתון כי כל השיפועים (b) במודל, שונים במובהק מאפס, ושערך β (הציון המתוקנן של b- השיפועים) של כל אחד מהמשתנים, שווה למתאם פירסון בין אותו משתנה לy. כלומר-ההשפעה הישירה של כל אחד מהמשתנים על y, לא יכולה להיות אפס. לפי הסכימה של רייט, המתאם הפשוט (פירסון) בין 2 משתנים הוא סכום ההשפעות הישירות והקשרים העקיפים, כלומר- שווה ל βבין X מסוים לY, ועוד כלל ההשפעות העקיפות על המשתנה (המתאם שלו מול כלx מנבא אחר, כפול ההשפעה הישירה של אותו ה xעל y). אם כך, ברגרסיה פשוטה, בטא פשוט שווה למתאם פירסון בין משתנה x לy (יש רק משתנה אחד- אין קשרים עקיפים). שמעון השתמש ברגרסיה מרובה, ומה שמסביר את מה שקיבל, זה שכנראה אין קשרים עקיפים בין האיקסים השונים. כשאין קשרים עקיפים (כל המתאמים בין המשתנים האיקסים המנבאים הם אפס) אז כל ערך בטא יהיה שווה למתאם עצמו (ממש כמו ברגרסיה פשוטה), היות שמעבר לאיבר בטא, כלל האיברים מתאפסים (כי כופלים אותם במתאמים שמייצגים את הקשרים העקיפים שלא קיימים במקרה זה), ונשארים רק עם זה שהמתאם שווה בטא.

**3.** **נכון.** SStotal=0 הוא סכום ריבועי הסטיות מהמודל (השונות הלא מוסברת) כאשר התבצע ניבוי באמצעות הממוצע. כאשר SStotal=0 זה אומר שכל השונות מוסברת על ידי מודל הממוצע, ולכן הוא המודל הטוב ביותר. גם אם היינו בונים קו הרגרסיה הוא היה על הממוצע (משום שאין שונות אז ה"משתנה" הוא בכלל קבוע, וכל ערכי המדגם שווים לממוצע). אם אין לי שונות להסביר, אין סיבה לבנות מודל רגרסיה על מנת לשפר את הניבוי.

**4. לא נכון.** בבניית מודל רגרסיה נרצה שהפער בין שני הR המרובעים יהיה כמה שיותר קטן (יעיד שככל הנראה אין התאמת יתר), אך ללא קשר להתאמת יתר, תמיד סביר כי הR בריבוע של סט האימון יהיה גדול (או שווה) לR בריבוע של סט הבחינה משום שהמודל פשוט הכי יתאים לנתונים עליהם הוא נבנה, וסט האימון גם מבוסס על רוב הנתונים. באופן כללי, נחשוד בהתאמת יתר אם יש פער מאוד גדול בין שני הסטים, ואם נתון לנו שמדגם עליו ערכו תיקוף צולב היה קטן (גם ביחס לכמות המנבאים במודל). אך כאן, לא נאמר לנו מה הפער בין שני הסטים ולא מובאים נתונים, לכן לא ניתן להסיק שהפער בין הRים בריבוע אכן מספיק גדול ולהכריע על התאמת יתר. **שאלות פתוחות**  
**א.** משוואת הרגרסיה של המודל של שירין, מנבאת את המשתנה המנובא (ממוצע הדירוגים), מתוך משתנה מנבא אחד בלבד (המידה בה הנבדקים הרגישו שרשימת הסרטים הותאמה אליהם אישית), לכן- זוהי משוואת רגרסיה לינארית פשוטה:

ניתן לכתוב גם כך:

לעומת זאת, עדי מציעה מודל שמשוואת הרגרסיה שלו מנבאת את המשתנה המנובא (ממוצע הדירוגים), מתוך שני משתנים מנבאים הקשורים באישיות של הנבדקים (Agreeableness ו-Openness ), לכן- זוהי משוואת רגרסיה לינארית מרובה:

ניתן לכתוב גם כך:

תמונה שמכילה טקסט, צילום מסך, מחשב, כמה

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי**ב.** על מנת לבחון האם השערת חוקרת אוששה, ראשית יש לוודא שמשוואת הרגרסיה שלה בכלל יכולה לנבא את ערך ה-Y המנובא (ממוצע דירוגים) של נבדק מסוים היטב. לשם כך- ערכי כל המקדמים של משוואת הרגרסיה: השיפועים (b) והחותך שלה (a) צריכים להיות מובהקים. אם כל הערכים מובהקים (מודל סביר) נסתכל על השיפוע (ערך הb) שקיבלנו ונבחן האם הוא תואם את כיווניות ההשערה של החוקרת. בנינו את 2 המודלים והרצנו פקודת 'summary' לכל מודל בנפרד, שנתנה לנו מידע על מובהקות המקדמים של 2 משוואות הרגרסיה. במודל של שירין (מודל 1) כל הערכים מובהקים, ולכן המודל שלה סביר. בנוסף- ניתן לראות שהשיפוע (b), הוא חיובי, מה שתואם את השערתה שהמידה שבה נבדקים הרגישו שרשימת הסרטים הותאמה אליהם אישית (Is\_personalized) תביא לדירוג סרטים ממוצע גבוה יותר, והשערתה אוששה. לעומת זאת במודל 2 (של עדי), ערך המקדם b של openness אינו שונה במובהק מאפס (p-value>0.05), אז למרות שערך המקדם של Agreeableness מובהק, וכיוונו מתאים להשערה, ולמרות שהמודל כולו מובהק, השערתה (שAgreeableness וOpenness ישפיעו) לא אוששה במלואה.

**ג.** באופן כללי, b הוא שיפוע (מקדם רגרסיה), המייצג את השפעתו של משתנה מסוים על משתנה מנובא y (כשאני עולה ביח' אחת בx כמה יח' עליתי בy). β הוא השיפוע המתוקנן (בכמה ס"ת ישתנה y עבור עליה בס"ת אחת בx), המהווה מדד להשפעה ישירה של משתנה. ברגרסיה פשוטה ה βשווה למתאם הפשוט (r) בין x וy (חותך תמיד 0). ברגרסיה מרובה יש לזכור שכל מתאם פשוט בין x מסוים לy מורכב מהשפעות עקיפות והשפעה ישירה, לכן ע"מ לבודד ולהשוות השפעות ישירות נשתמש בערכי β מוחלטים (בציוני תקן יחידות מדידה מנוטרלות, בעוד שערכי הb מלוכלכים ובסולמות מדידה שונים). במודל של שירין 0.08b=: עליה ביח' אחת בis\_personalized (1-5) תביא לעליה ב0.08 יח' בavg\_rating (0.5-5 כוכבים). ו:β=0.21 עליה בס"ת אחת בis\_personalized תביא לעליה ב0.21 ס"ת בavg\_rating. במודל של עדי b1 לא שונה במובהק מ0 לכן לא תורם (וכך גם β1). ו0.06b2= אז על עליה ביח' אחת ב agreeableness(1-7) יש עליה ב0.06 יח' בavg\_rating כשהמשתנה openness מוחזק קבוע.ו0.17 β2=אז עליה בס"ת אחת בagreeableness, תביא לעליה ב0.17 ס"ת בavg\_rating, כשהמשתנה openness מוחזק קבוע.

תמונה שמכילה טקסט, צילום מסך, מחשב

התיאור נוצר באופן אוטומטי

|  |  |
| --- | --- |
| **מודל 2- עדי** | **משמעות** |
| 0.03561 b1= | על כל עליה ביחידה אחת בopenness של נבדק (1-7), יש עליה ב0.03561 יחידות בדירוג הסרטים הממוצע שלו, כאשר המשתנה agreeableness מוחזק קבוע. |
| 0.06360 b2= | על כל עליה ביחידה אחת ב agreeablenessשל נבדק (1-7), יש עליה ב0.06360 יחידות בדירוג הסרטים הממוצע שלו, כאשר המשתנה openness מוחזק קבוע. |
| 0.08 β1= | עליה בסטיית תקן אחת בopenness של נבדק, תביא לעליה ב0.08 סטיות תקן בדירוג הסרטים הממוצע שלו, כאשר המשתנה agreeableness מוחזק קבוע. |
| 0.17 β2= | עליה בסטיית תקן אחת ב agreeablenessשל נבדק, תביא לעליה ב0.17 סטיות תקן בדירוג הסרטים הממוצע שלו, כאשר המשתנה openness מוחזק קבוע. |

|  |  |
| --- | --- |
| **מודל 1- שירין** | **משמעות** |
| 0.08181 b= | עליה ביחידה אחת של המידה בה נבדק הרגיש שרשימת הסרטים הותאמה לו אישית (1-5), תביא לעליה של 0.08181 יחידות בדירוג הסרטים הממוצע שלו (0.5-5 כוכבים). |
| 0.21 β= | עליה בסטיית תקן אחת של המידה בה נבדק הרגיש שרשימת הסרטים הותאמה לו אישית, תביא לעליה של 0.21 סטיות תקן בדירוג הסרטים הממוצע של הנבדק. |

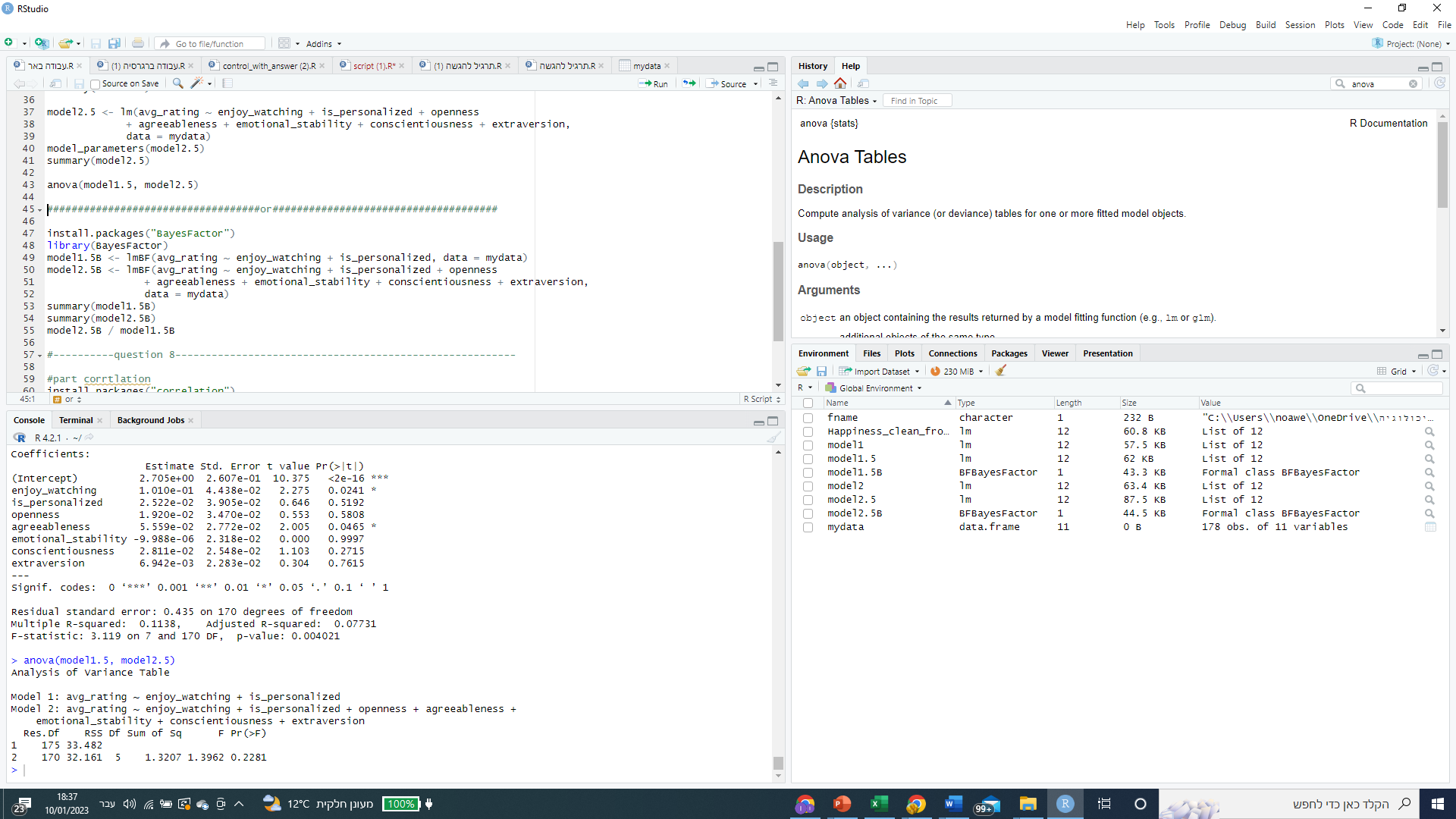
\*בסעיפים א' וב' חילצנו את השיפועים (b) של שתי משוואות הרגרסיה. את ערכי ה βניתן להוציא מקוד סעיף זה.  
\*\*מצורפות טבלאות מסודרות עם הנתונים כנספח (המידע גם קיים בשאלה, אך כאן זה יותר מסודר):

**ד.** סכימה של רייט מייצגת פרשנות אפשרית אחת לדפוס קשרים בין משתנים, תחת מודל סיבתי תאורטי (והנחת לינאריות). לפיה בין כל 2 משתנים יש קשרים שמושפעים מהרבה נתיבים, חלקם ישירים (השפעות ישירות), וחלק עקיפים (קשרים עקיפים), ומתאם פשוט בין X לY הוא סכום כל הנתיבים הללו. כשנרצה למצוא תרומה ייחודית של משתנה עלינו להתחשב במשתנים מתערבים, ולנטרל את הקשרים העקיפים בכך שנחזיק אותם קבועים. רועי לא בהכרח צודק. אולם יש סיכוי שבעולם הקשר בין יציבות רגשית לממוצעי דירוגי הסרטים חלש, ולכן המשתנה באמת לא יתרום לניבוי במודל הרגרסיה, אך לפי רייט, חוזק הקשרים בין מנבאים משפיע על המודל ועל משוואת רגרסיה מרובה, ומתאם פשוט לא יעיד לי על כמות השונות המוסברת הנקייה שמשתנה יכול לתת לי. ייתכן מצב של דיכוי, בו שליטה על משתנה מתערב מחזקת אפקט של משתנה עניין (x). הכנסתו של משתנה מדכא תביא לכך ששיפוע משתנה העניין יתחזק/ יהפוך סימן (הקשר מתחזק בכיוון הנכון). בעצם, בהכנסת המשתנה יציבות רגשית למודל, בצירוף עם השפעות עקיפות רלוונטיות, ייתכן שנזהה שהשפעתו הישירה גדולה משחשבנו/שהשפעה ישירה של משתנה אחר גדולה משחשבנו (כתלות במי משתנה מדכא), וכך הוספת המשתנה גם תגדיל את אחוז השונות המוסברת של המודל.

**ה.** על נופר להשתמש ברגרסיה היררכית (השוואה בין מודלים כאשר אחד מקונן בשני), בה נשתמש בכדי להכניס קבוצות של משתנים ולראות את התרומה הכוללת שלהם למודל. כדי לבחון את השפעת בלוק הbig five, על נופר ליצור 2 מודלים: מודל הבסיס- מנבא את התרומה הייחודית של 2 משתני הרקע כקבוצה (במקרה זה המנבאים: is\_personalized ו-enjoy\_watching), ומודל בסיס+ כל משתני הbig five (כמנבאים גם כן). לאחר מכן, תוכל לבצע מבחן F למובהקות התוספת שיענה על השאלה האם השינוי (הגדלה) בR בריבוע בעקבות הוספת המשתנים הוא מובהק. תוצאה מובהקת במבחן F, תעיד שהוספת המנבאים למודל הגדילה את אחוז השונות המוסברת על ידו בצורה מובהקת, ותאפשר להכריע כי זהו מודל עדיף, שהמשתנים הנוספים בו תורמים מעבר למשתני הבסיס. את מובהקות התוספת של משתני ה big five נבדוק באמצעות מבחן ANOVA המשווה בין 2 המודלים. נמצא כי התוספת של הR בריבוע במודל המורכב בהשוואה למודל הבסיס לא מובהקת (p-value>0.05), אז תוספת משתני הbig five לא תרמה לשונות המוסברת במודל מעבר למשתנים: is\_personalized ו-enjoy\_watching, ולא נדחה את השערת האפס שאומרת שמודל הבסיס הוא העדיף.

תמונה שמכילה טקסט, צילום מסך, מחשב נישא

התיאור נוצר באופן אוטומטי

\* הערה: אם נרצה לקבל את השערת האפס ולהכריע שמודל הבסיס עדיף, אפשר גם להשתמש בשיטה בייסיאנית.

**ו.** R בריבוע נמוך יותר במודל החדש של דביר, בהשוואה למודל עם כל ה-Big Five לא בהכרח יכול להוות עדות נגדית לטענתו של דביר שמודל בלי Conscientiousness הוא עדיף. הסיבה לכך היא שהוספת כל משתנה שהוא למודל רגרסיה מרובה (גם אם מנבא היטב וגם אם לא), תגרום לR בריבוע לעלות/ לא להשתנות, ובהתאמה (עובד באותו אופן גם הפוך) הורדת כל משתנה שהוא מהמודל, תוריד/ לא תשנה את הערך שלו. כלומר, הירידה שנצפתה בהשוואת המודלים הגיונית ועשויה לקרות בלי קשר לכמה מודל עדיף, וזה לא אומר שזה הופך את המודל לפחות טוב (לעיתים החסרת משתנה מנבא אף יכולה למנוע over-fitting ולייצר מודל משופר). על מנת להבין מי המודל העדיף, יש לבדוק עד כמה המנבא שהוסף באמת תרם להסברת אחוז השונות המוסברת. בעצם, נרצה לבדוק האם ההקטנה של SSres במודל החדש (אחוז השונות הלא מוסברת שבהוספת מנבא תמיד תקטן/תשאר זהה) מצדיקה את הוספת המנבא. דביר יכול לבדוק זאת ב2 שיטות: בייסיאנית (בייס פקטור)/ שכיחותנית (מבחן מובהקות).

**ז.** במצב המתואר, נציע לדיסמי להשתמש במדד מרכז כמודל כתלות בפונקציית ההפסד (נרצה שתהיה קטנה ככל הניתן), ובאופן ספציפי- בשכיח כמודל. בבניית מודל סטטיסטי נתבסס על נתונים קיימים, ונשתמש בידע זה כדי לנבא נתונים/תצפיות חדשים לגבי העתיד. בחירה במדד מרכז נובעת מכך שהנתונים היחידים שבידנו הם דירוגי מאה צופים שצפו בסרט החדש של דיסמי ודירגו אותו בסולם של 1-5 כוכבים, ואין לנו שום דרך לנבא את הדירוג שלהם ממשתנים מסוימים, שכן רק Y נתון לנו עבור כל צופה. פונקציית הפסד היא מדד לכישלון המודל שנרצה שתהיה הקטנה ביותר, לכן עלינו לבחור במדד מרכז התואם את פונקציית ההפסד בה חברת דיסמי משתמשת לבחירת המודל הזוכה. לפיהם, המודל הזוכה הוא זה שניבא את הדירוג המדויק הכי הרבה פעמים (למשל, ניבוי מדויק הוא שכאשר נבדק מדרג את הסרט ב3 כוכבים, המודל ינבא שהנבדק ידרג את הסרט ב3 כוכבים). כלומר-דיסמי סופרים מספר הצלחות, ולכן פונקציית ההפסד בה הם משתמשים היא מספר הטעויות. עבור פונקציית הפסד זו מדד המרכז בו הכי כדאי להשתמש הוא השכיח (ייתן פונקציית הפסד מינימלית) ולכן נשתמש בו כמודל.

**ח.** ניתן לערוך בקרה סטטיסטית באמצעות מתאם חלק, שהוא המתאם בין מנבא מסוים למשתנה המנובא (Y), בניכוי השפעת משתנה נוסף מאחד מהם. במקרה זה, מדובר על הקשר בין enjoy\_watching לis\_personalized, כמלכלך את הקשר שלenjoy\_watching עם avg\_rating , ולכן יש לבחון את הקשר של הציפייה לגבי הנאה על ממוצע דירוגי הסרטים, בניכוי המידה בה נבדקים חשו שהרשימה מותאמת אישית להם. לשם כך, נחשב מתאם חלק (נקי יותר) בין ציפייה להנאה בניכוי is\_personalized, לavg\_rating. תחילה, ננבא את enjoy\_watching מis\_personalized. לאחר הניבוי, ניצור את וקטור הרסידואלים (SSres החלק ב enjoy\_watchingשלא מוסבר ע"י is\_personalized), על ידי חיסור כל ציון מנובא של נבדק מאמיתי שלו. נחשב את המתאם בין ציוני ההפרש לבין avg\_rating, על מנת לקבל את מתאם החלק ביניהם. בR יש קוד שמפשט את זה, וגם שם מתאם החלק יצא 0.18 (נמוך מהמתאם הפשוט 0.28 כצפוי) ומובהק (p-value<0.05), לכן- דימה צדק שיש קשר בין avg\_rating לenjoy\_watching, שלא נובע מחפיפה אפשרית שלו עם is\_personalized (השפעתו "מלכלכת" את המתאם הפשוט בין enjoy\_watching וavg\_rating).

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי