דוח הפרויקט

מגישים: עדריאל מנדלסון ומרים בינהורן

מבוא:

כסטודנטים בעצמנו, התעניינו לבדוק כיצד כלל הנתונים לגבי אורח החיים של הסטודנטים שנבדקו השפיע עליהם, ובפרט שאלת המחקר שלנו הייתה: **אילו גורמים הם המשפיעים ביותר על איכות חייהם של סטודנטים ועל הציונים שלהם, ובפרט איזה תפקיד השינה משחקת בזה?**

הדוח שלפניכם עוסק בניתוח הקשר בין אורח חיי הסטודנטים לבין איכות חייהם וממוצע ציוניהם. במסגרת הפרויקט, בחנו נתונים מתוך שני מאגרי מידע, שכללו מידע על מדדים מגוונים, כמו רמת השינה, רמות הלחץ, הפעילות הגופנית ועוד. שאלת המחקר המרכזית הייתה לזהות מהם הגורמים המשפיעים ביותר על איכות החיים וההצלחה האקדמית של הסטודנטים, עם דגש מיוחד על תפקידה של השינה.
תהליך העבודה כלל התמודדות עם מספר אתגרים, כגון חילוץ מספר גדול של נתונים מפורמט לא מוכר - קבצי JSON, הבנת מדדים מורכבים, וטיפול בערכים חסרים. לאחר בחירת המשתנים הרלוונטיים, השתמשנו בטכניקות ניתוח שונות, כמו יצירת טבלאות קורלציה, חישוב רגרסיות לינאריות ומפות חום. הדוח מציג את המסקנות שעלו מהנתונים הראשוניים ומהעמקה במאגר נוסף שהתמקד ספציפית בקשר שבין שינה לציונים.

תהליך העבודה:

הדאטה שאיתו עבדנו מורכב מהמון נתונים כלליים לגבי חייהם של סטודנטים שנבדקו במחקר ונעקבו במספר אמצעים שונים. הנתונים העיקריים שבהם בחרנו להתמקד הם שאלונים שהסטודנטים מילאו באופן עצמאי ומידע לגבי ממוצע הציונים שלהם.

האתגר הראשון שלנו בניסיון למצוא מענה לשאלה זו היה חילוץ הנתונים מהdata-set שמצאנו:

- ראשית, רוב הקבצים היו שמורים בפורמט json בו לא נתקלנו מעולם לפני כן, והיינו צריכים ללמוד להבין אותו ואילו כלים עומדים לרשותנו כדי לנתח אותו בעזרת פייתון בצורה יעילה.

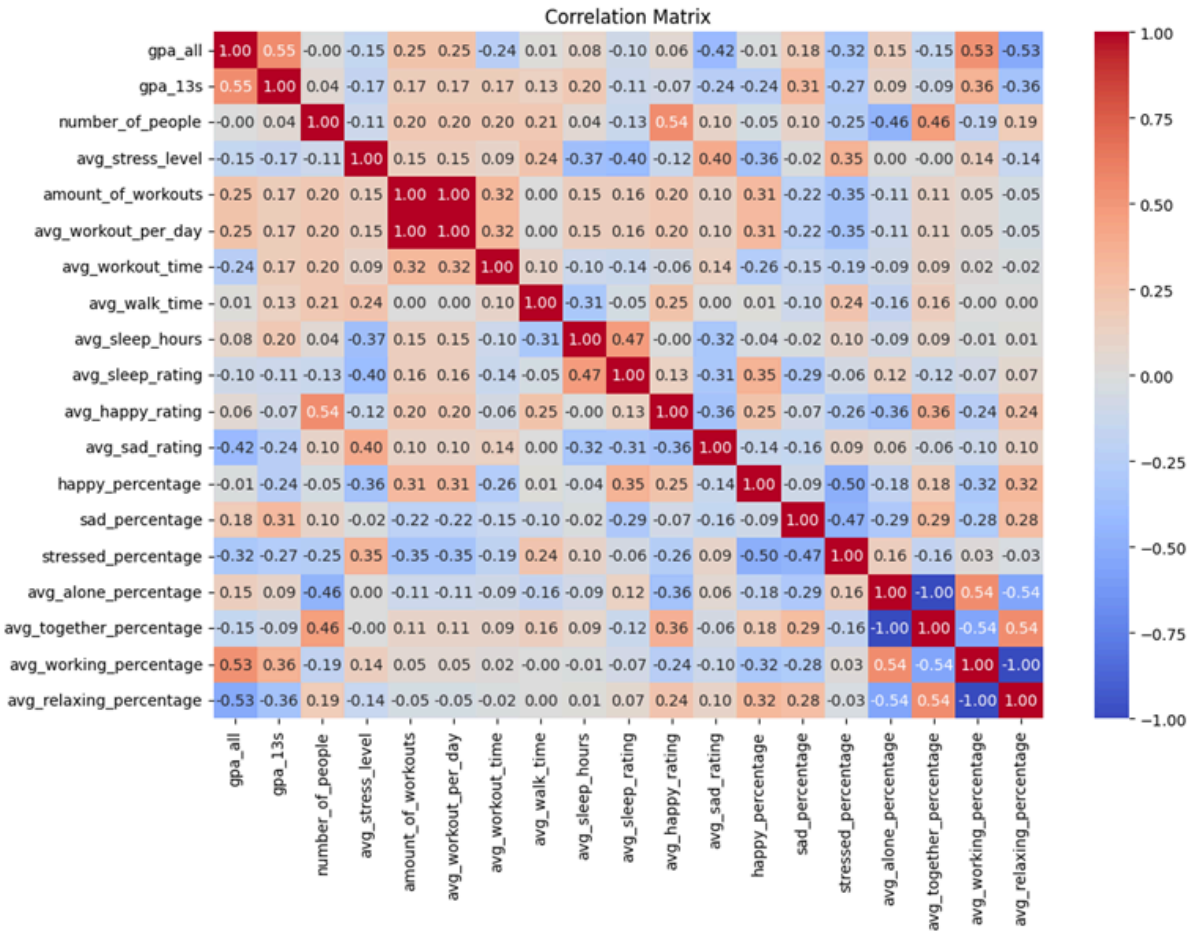
- שנית, המדדים בהם השתמשו החוקרים ברוב השאלונים לא היו אינטואיטיביים (כיוון שרוב השאלות היו אמריקאיות, אך הכילו נתונים מורכבים לייצוג בצורה זו), ולכן נאלצנו להתאמץ הרבה כדי להבין את הנתונים בצורה נכונה, ולאחר מכן לכתוב קוד שימיר את הערכים לכאלו שמתאימים לחישובים שאותם אנחנו תיכננו לעשות.

- טיפול בערכים חסרים – הנתונים הכילו סטודנטים שחלק מהמידע או כולו לא היה נתון עבורם. היינו צריכים להבין איך להתייחס אליהם וכיצד לבצע את ניתוח הנתונים כך שיתבסס על כמה שיותר מידע אבל בו-זמנית שיתבסס רק על דאטה אמין ולא כזה שנוצר ממחסור בנתונים אמיתיים.

בגלל כל המגבלות האלו והעובדה שהדאטה שמצאנו היה עצום ורצינו להתמקד רק בחלק ממנו, התחלנו בלבחור נתונים שאיתם אנחנו רוצים לעבוד, וחילצנו אותם לתוך טבלת אקסל שאיתה יהיה לנו קל יותר לעבוד. כך למשל במקום להתעסק עם כל הדיווחים של כל סטודנט לגבי כמה שעות הוא ישן בכל לילה, הזנו עבור כל סטודנט רק את כמות שעות השינה הממוצעת שלו בלילה, ועל זה ביצענו בהמשך את ניתוח הנתונים שלנו.

לאחר שהעברנו את המידע לטבלה, התחלנו סוף-סוף בביצוע אנליזות. ראשית רצינו לראות בין אילו היבטים בחיים של הסטודנטים יש קורלציות גבוהות יחסית, ולכן יצרנו טבלה שמחשבת ומציגה את הקורלציה בין כל זוג משתנים. הנה התוצאה שקיבלנו:

*(*** פירוט משמעות המשתנים מופיע בסוף המסמך)*



נרצה לשים לב לכמה דברים, שמסמלים לנו שהתוצאה ככל הנראה אמינה – קיבלנו אלכסון שכל הערכים בו הם 1, כיוון שלכל משתנה יש כמובן התאמה מושלמת עם עצמו.

יש קורלציה חיובית גבוהה בין gpa_all לבין gpa_13s, שמסמלים בהתאמה את ממוצע הציונים של הסטודנט עד כה בתואר ואת ממוצע הציונים שלו בסמסטר הספציפי בו הוא נבדק בניסוי.

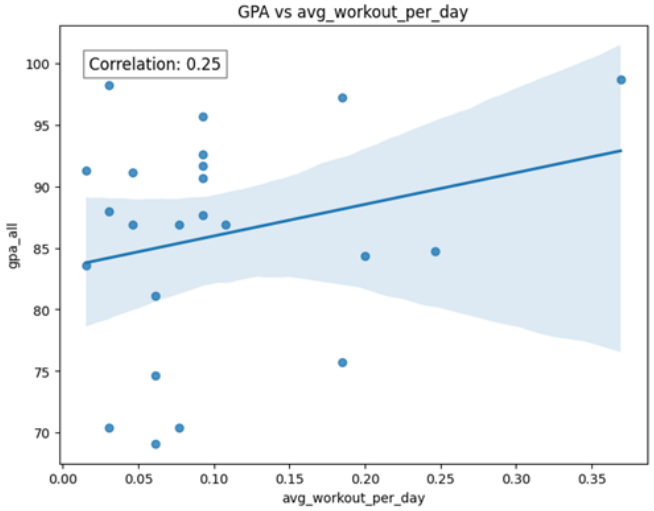
עוד קיבלנו קורלציה חיובית גבוהה בין כמות שעות שינה לבין איכות השינה שדירג הסטודנט, בין כמות האנשים איתם הוא נפגש לרמת האושר שלו ובין הזמן שהוא משקיע בלמידה לבין ממוצע הציונים שלו.

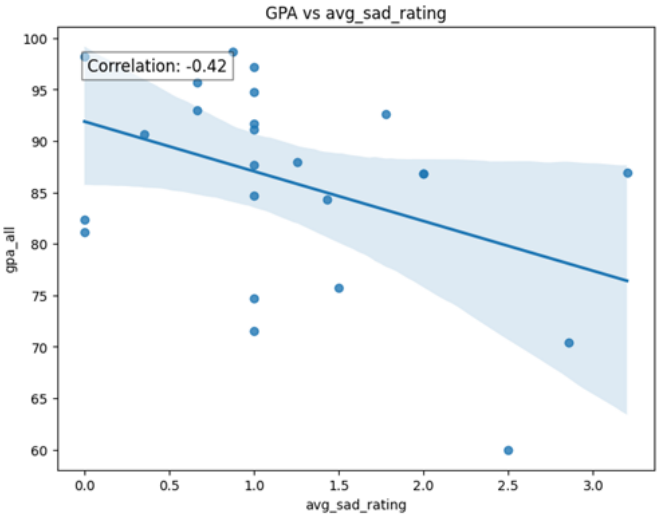
קורלציה שלילית לעומת זאת ניתן לראות בין רמת הלחץ של הסטודנט לאיכות השינה שלו, ובין התעמלות גופנית ללחץ.

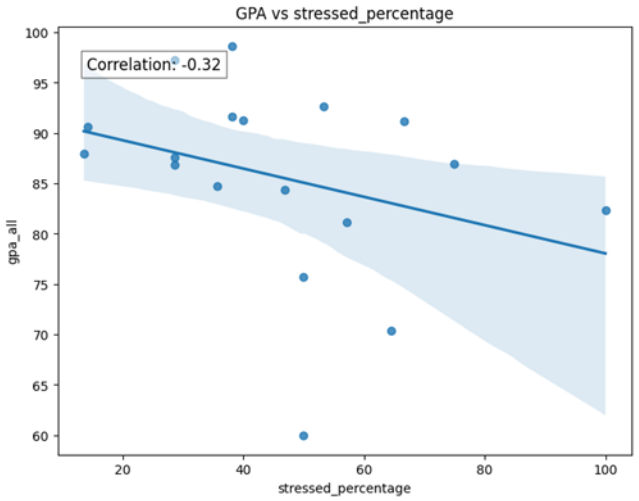
כל אלו נראים לנו כמו מידע אמין שכן הוא תאם לציפיות שלנו, ולכן המשכנו הלאה לניתוחים נוספים ומעמיקים יותר בנתונים שנראה לנו כי הקורלציה ביניהם תניב תוצאות משמעותיות יותר, תוך התמקדות בנושאים שבהם עוסקת שאלת המחקר שלנו שהם שינה וממוצע ציונים. להלן כמה מהתוצאות שקיבלנו:

*(*** דגש לגבי כלל הגרפים המובאים כאן: הקורלציה בצד מייצגת את ערך מקדם הקורלציה של פירסון עבור הגרף הנתון. הקו הכחול הוא קו הרגרסיה הלינארית, והאזור התחום בתכלת מסביבו הוא "רצועת האמון" שמייצג את טווח האי-ודאות של הרגרסיה – כלומר כאשר הרצועה צרה, זה אומר שהמודל בטוח יותר בתחזיותיו לגבי הקשר בין המשתנים, וככל שהיא רחבה יותר כך המידע פחות אמין.)*

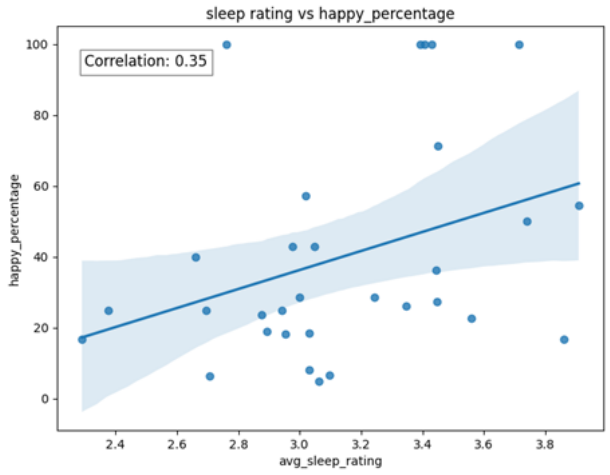
הגורמים המשפיעים ביותר על ממוצע הציונים על פי הנתונים הם: כמות זמן הלמידה (בקורלציה חיובית), רמת הלחץ (בקורלציה שלילית), עד כמה הסטודנט חש עצוב (בקורלציה שלילית), כמות התעמלות גופנית (בקורלציה חיובית). להלן הגרפים הממחישים זאת:

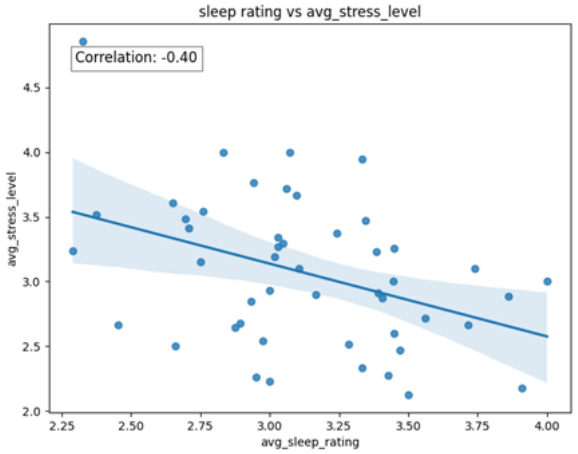






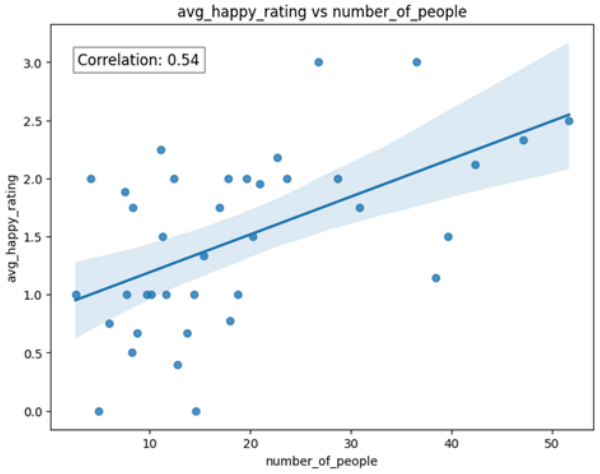
לעומת זאת, הגורמים שהתגלו כמשפיעים ביותר על איכות השינה של הסטודנט הם דווקא: רמת הלחץ שלו (קורלציה שלילית) ואחוז הזמן בו הוא חש מאושר (קורלציה חיובית). להלן הגרפים:



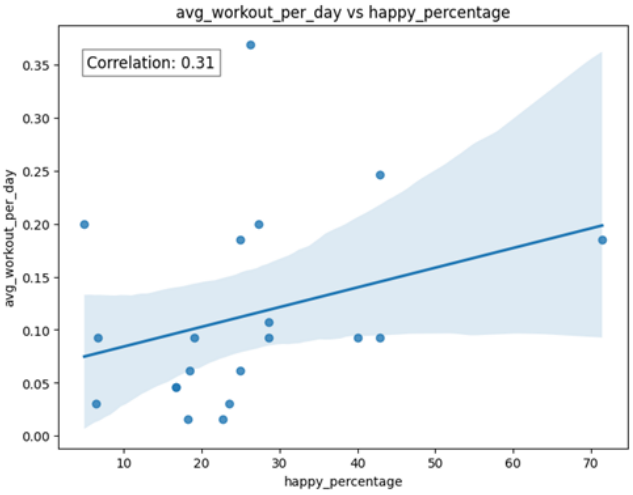


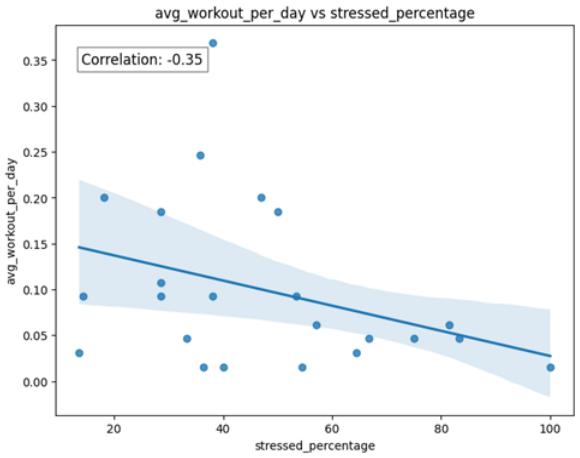
בנוסף, מצאנו כי מהנתונים עולות מגוון תוצאות מעניינות נוספות:

קורלציה חיובית גבוהה בין דירוג האושר לבין כמות האנשים איתם הסטודנט נפגש ביום

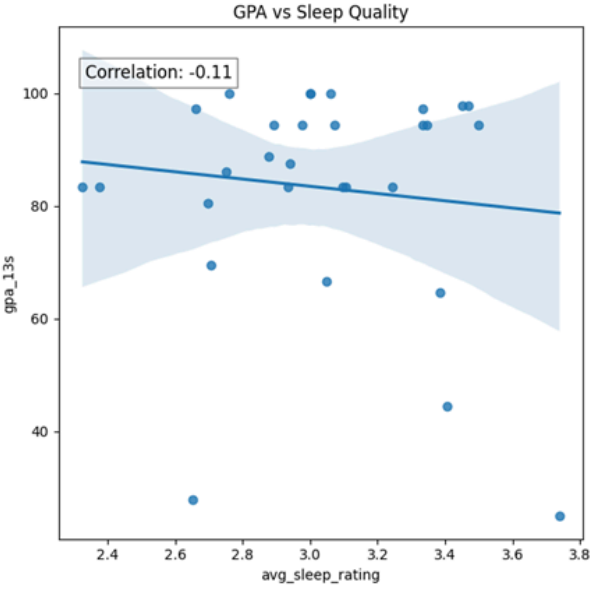


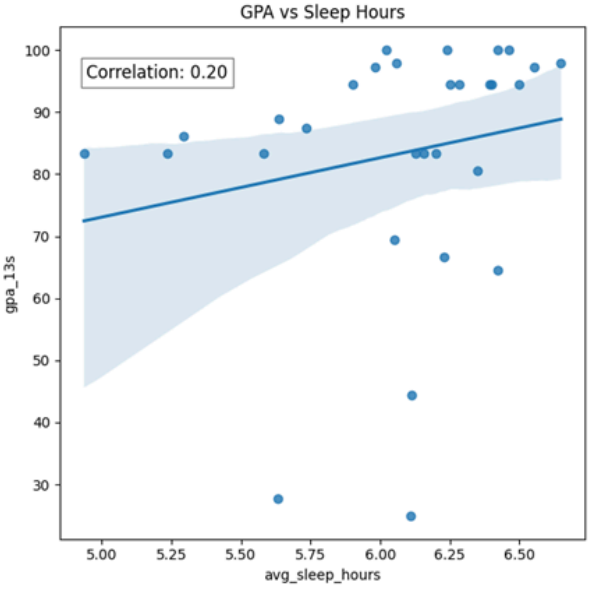
מבין שלושת הגורמים לחץ, שמחה ועצב, ההשפעה העיקרית של התעמלות גופנית היא הפחתת הלחץ והעלאת השמחה:





כל זה גילה לנו כבר הרבה לגבי הגורמים המשפיעים על אורח חייהם של הסטודנטים, אך בין 2 הגורמים בהם שאלת המחקר שלנו התמקדה – שינה וממוצע ציונים – לא נראה כי יש בנתונים שלנו קורלציה משמעותית:



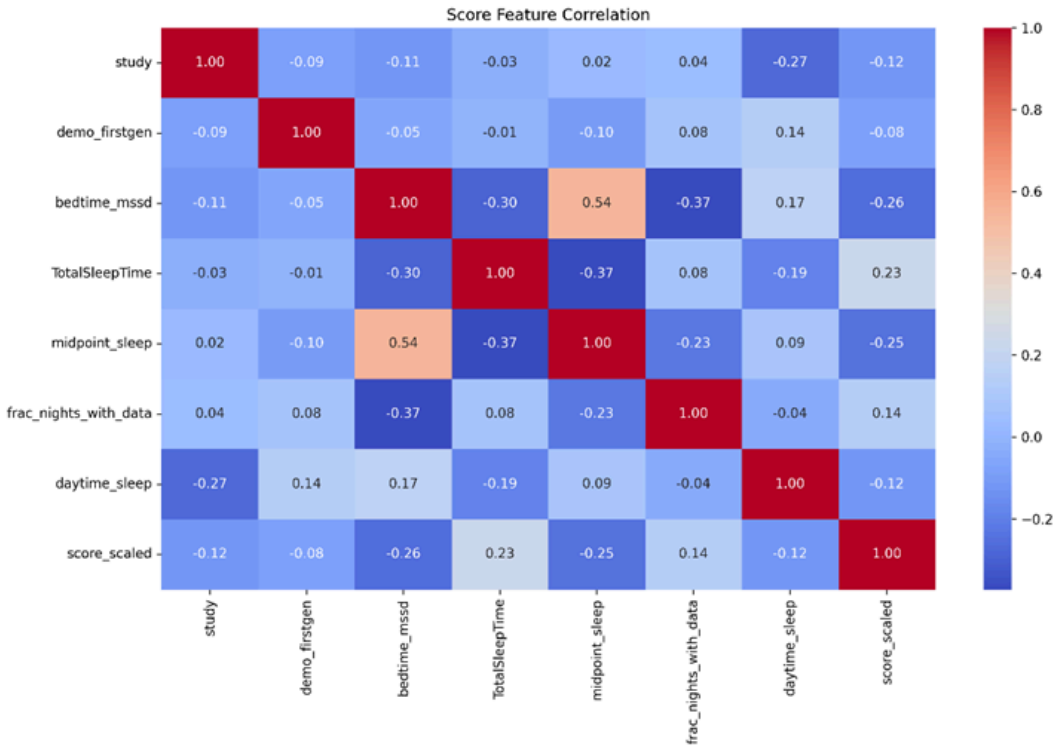


כפי שניתן לראות, מקדמי פירסון נמוכים למדי, והופתענו לראות כי בעוד הקורלציה בין ממוצע ציונים לשעות שינה היא חיובית, דווקא הקורלציה בין איכות השינה לממוצע היא שלילית. בנוסף רצועת האמון ברובה מאוד רחבה, כלומר שהקורלציות האלו לא מבוססות על מספיק נתונים.

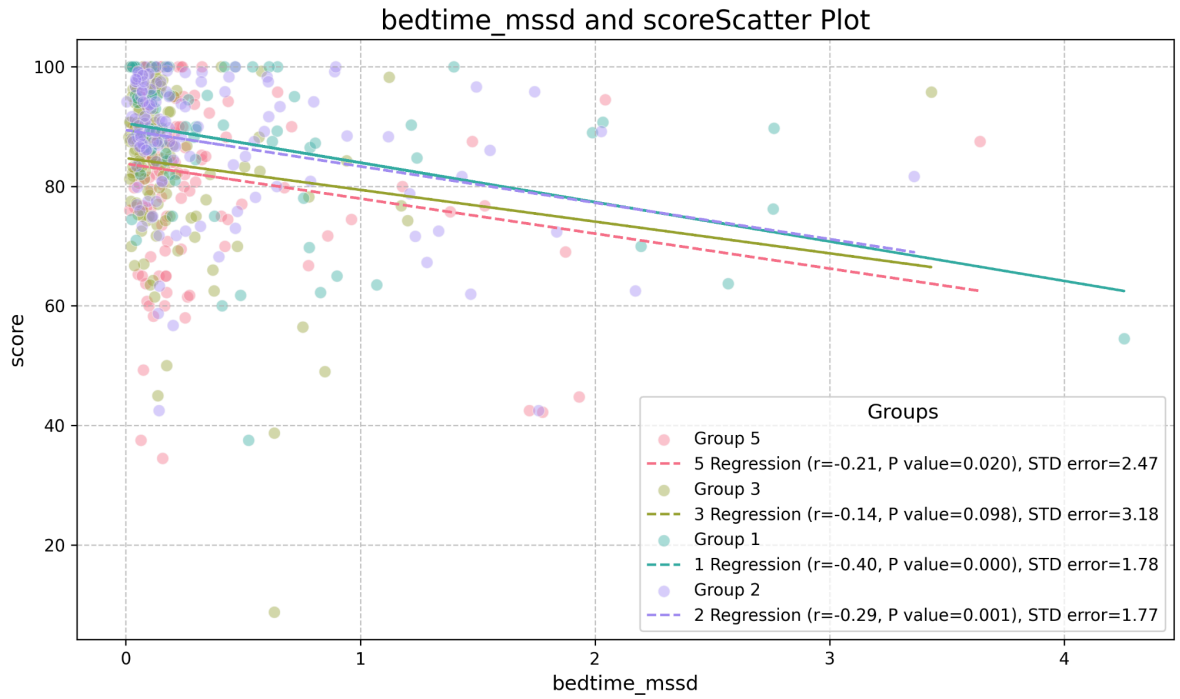
בהתחשב בכל זה, החלטנו לפנות למסד נתונים נוסף שעוסק רק בנושאים אלו – שינה וממוצע ציונים. כלומר, מהמסד הראשון כבר למדנו הרבה לגבי מגוון נושאים הקשורים לאורח החיים שלנו כסטודנטים, אך לכך נוסיף כעת מידע ממקור נוסף ששינה וציונים הם הנושאים היחידים שנבדקים בו ובאופן מאוד רחב (מעל 600 נבדקים).

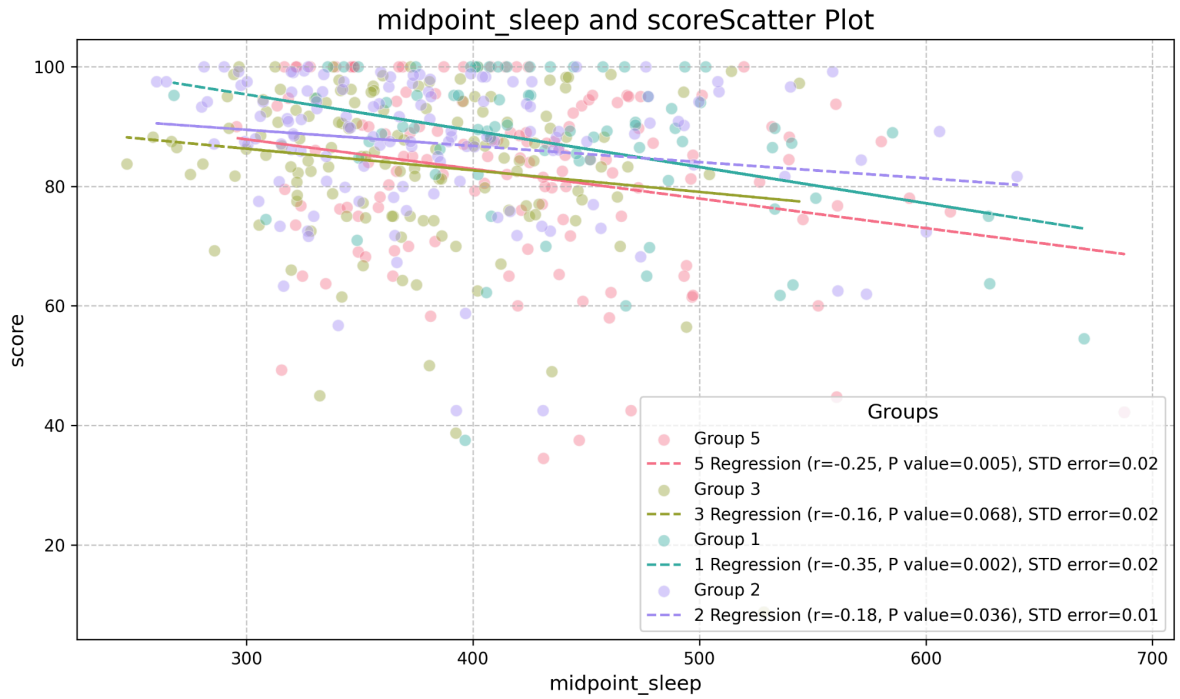
גם בdata set זה התחלנו ביצירת מפת חום לזיהוי קורלציות בין כלל המשתנים הרלוונטיים:

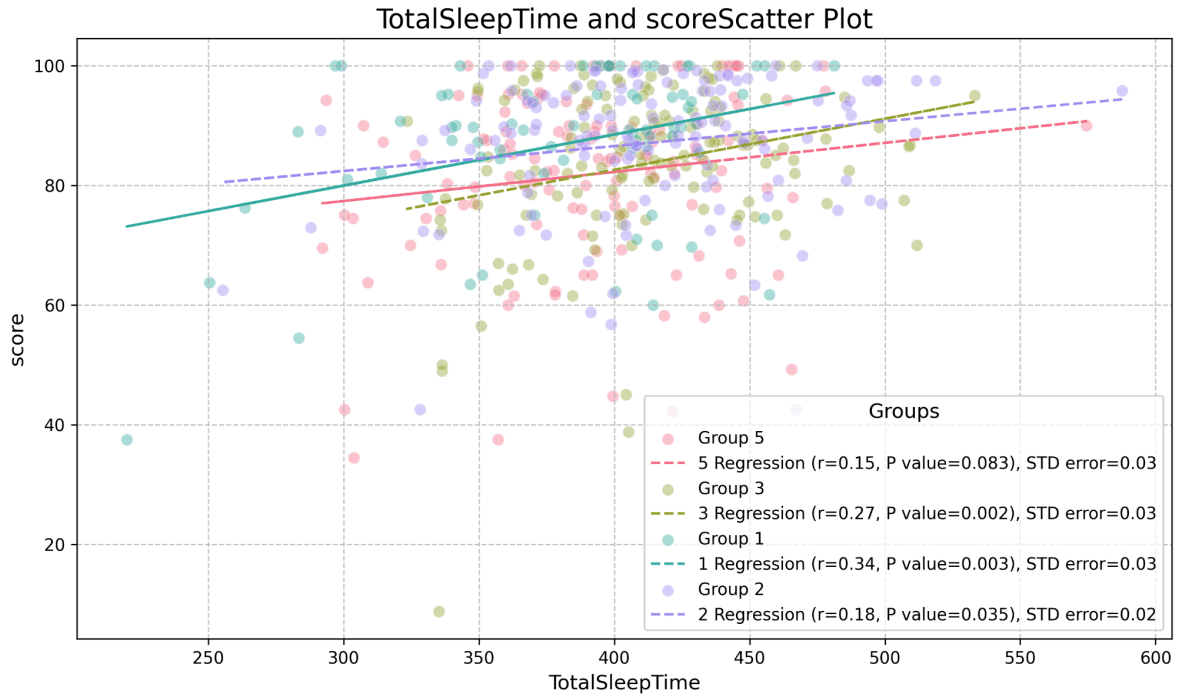
*(*** פירוט משמעות המשתנים מופיע בסוף המסמך)*



לאחר מכן, רצינו לראות את הקשר בין הציונים לבין אספקטים שונים של שינה. בגרפים הבאים נציג רגרסיה לינארית בין ציון לבין השונות בשינה (עד כמה השינה של הסטודנט לא סדירה) , כמות שינה ממוצעת במהלך שעות היום (בדקות), וכמות שינה בממוצע בלילה (בדקות).







מסקנות:

ממסד הנתונים הראשון הסקנו כי הגורם המשפיע ביותר על ממוצע הציונים של סטודנט הוא כמות הזמן שהסטודנט מקדיש ללמידה. גורמים נוספים כוללים מצב רוח והתעמלות- סביר שממוצע הציונים של הסטודנט יעלה ככל שיהיה פחות עצוב ופחות לחוץ. גם התעמלות גופנית משפיעה לחיוב

על ממוצע הציונים, אך הוודאות בנוגע לכך נמוכה יותר (שכן המידע בנושא מבוסס על פחות נתונים ולכן פחות אמין).

לגבי גורמים אחרים הקשורים לאיכות החיים של הסטודנטים – מצאנו כי ככל שאיכות השינה טובה יותר כך סביר שהסטודנט יהיה יותר שמח ופחות לחוץ. כמו כן הוא יהיה שמח יותר ככל שיפגוש יותר אנשים במהלך היום, וככל שיתאמן יותר כך יהיה יותר שמח ופחות לחוץ.

באמצעות מסד הנתונים השני העמקנו את ההבנה שלנו לגבי הקשר בין שינה לבין ממוצע ציונים. גילינו ששינה לא סדירה ונמנום במהלך היום יובילו לפגיעה בציוני הסטודנט, אך שנת לילה ארוכה תוביל בסבירות גבוהה לעלייה בציונים שלו.

פיצ'ר נוסף:

לאחר ניתוח כל הנתונים, החלטנו להוסיף פיצ'ר נוסף מבוסס למידת מכונה ומנסה ליישם את התובנות העולות מהנתונים אודות הקשר בין שינה לציונים.

זהו בעצם מודל ששואל את המשתמש מספר שאלות לגבי הרגלי השינה שלו:

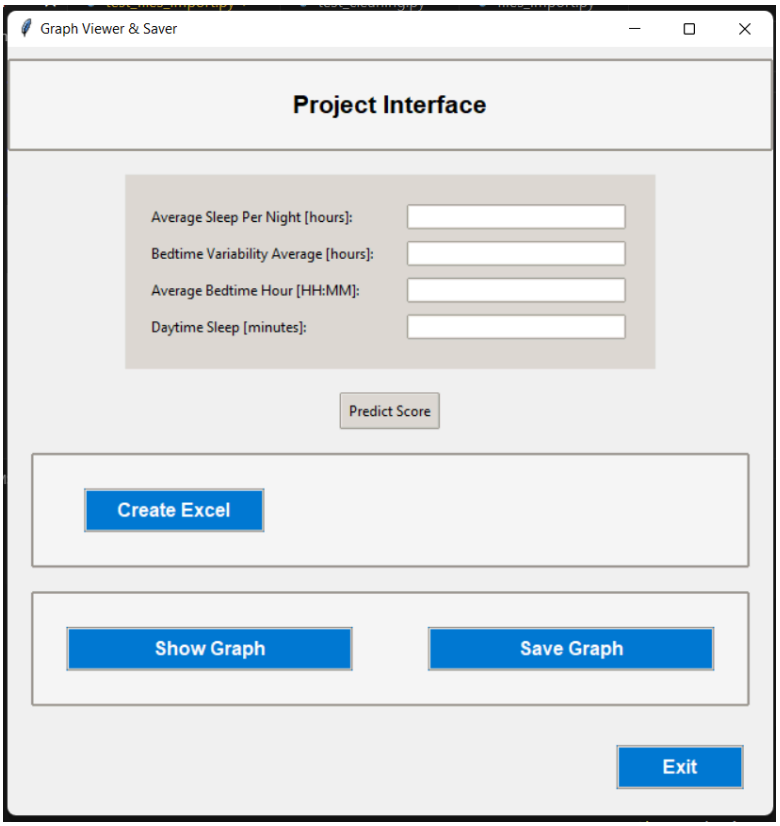
- Average Sleep Per Night [hours] - כמה שעות שינה אתה ישן בלילה בממוצע
- Bedtime Variability Average [hours] - מהי השונות בשעת השינה, כלומר מהו גודל הטווח בתוכו משתנה בדר"כ השעה בה אתה הולך לישון בלילה, בשעות
- Average Bedtime Hour [HH:MM] - באיזו שעה אתה הולך לישון בלילה ממוצע
- Daytime Sleep [minutes] - כמה דקות בממוצע אתה מנמנם במהלך היום

על סמך השאלות האלו והנתונים מהdataset השני שעוסק בקשר בין שינה לציונים, המודל מחשב את הציון הממוצע שהוא חוזה לאותו משתמש.
יש לציין כי זה כמובן לא מהווה מדד מדויק. השאלות עוסקות בשינה בלבד ועל פי הנתונים שראינו בdataset הראשון ישנם גורמים נוספים רבים שיש להם השפעה על ממוצע הציונים של סטודנט.
המטרה ביצירת המודל הייתה סה"כ המחשה של הנתונים שקיבלנו על הקשר בין שינה וממוצע ציונים. לגבי המידע הזה היה לנו דאטה הרבה יותר גדול לעבוד איתו וזה מה שהפך את יצירת המודל לאפשרית, ולכן הוספנו אותו על אף שהחישוב כרגע מתעלם מכל יתר ההיבטים של אורח חיי הסטודנט.

הרצת התוכנית:

בהרצת הפרויקט נפתח למשתמש חלון אינטראקטיבי בו עליו לבחור מה הוא רוצה:
הדבר הראשון שמופיע הוא המודל שבהזנת מענה לשאלות בו מציג למשתמש בחלונית חדשה את הציון הממוצע הצפוי לו על סמך הנתונים הללו.
לאחר מכן יש באפשרות המשתמש לבחור ב"Create Excel" בשביל להריץ את הקוד שכתבנו לחילוץ הדאטה המקורי מהdaraset הראשון ולהעבירו לטבלת אקסל עם העמודות שבחרנו להגדיר.
יש לשים לב כי במידה והדאטה לא קיים בחירה באפשרות זו תעלה הודעת שגיאה. (לגיטהאב לא העלנו את כל הדאטה הזאת כי היא מאוד כבדה, במידה והמשתמש ירצה להוריד אותה וליצור את הטבלה בה הניתוחים בהמשך משתמשים הוא יכול לעשות זאת.)

האפשרויות הבאות הן לקבל את הגרפים שאנחנו יצרנו להצגת הנתונים שעניינו אותנו, וגם שם יש 2 אפשרויות לבחירה - אם להציג את הגרפים (כל אחד מהם יפתח בחלונית חדשה) או לשמור אותם למחשב (הם ישמרו בתיקייה חדשה בשם results בתוך תיקיית הפרויקט).



נספחים

פירוט המשתנים באקסל של הdata-set הראשון:

User – המזהה הייחודי של הסטודנט

gpa_all – ממוצע הציונים הכללי של הסטודנט

gpa_13s – ממוצע הציונים של הסטודנט באותו סמסטר בו בוצע הניסוי

number_of_people – כמות האנשים שאיתם הסטודנט נפגש ביום בממוצע

avg_stress_level – רמת הלחץ הממוצעת של הסטודנט, בסולם של בין 1 ל5

amount_of_workouts – כמות הפעמים שהסטודנט התאמן בסמסטר

avg_workout_per_day – כמה זמן של אימון ביצע הסטודנט ביום בממוצע

avg_workout_time avg – מה היה אורכו של אימון בממוצע

walk_time – כמה זמן בממוצע הלך הסטודנט ביום

avg_sleep_hours – כמות שעות שינה ממוצעת בלילה

avg_sleep_rating – דירוג השינה הממוצע של הסטודנט, בסולם של בין 1 ל4

avg_happy_rating – דירוג ממוצע של רמת השמחה של הסטודנט, בסולם של בין 1 ל4

avg_sad_rating – דירוג ממוצע של רמת העצב של הסטודנט, בסולם של בין 1 ל4

happy_percentage – אחוז הפעמים בהם התשובה לשאלה מה התחושה הכי משמעותית שהסטודנט חווה הייתה שמחה

sad_percentage – אחוז הפעמים בהם התשובה לשאלה מה התחושה הכי משמעותית שהסטודנט חווה הייתה עצב

stressed_percentage – אחוז הפעמים בהם התשובה לשאלה מה התחושה הכי משמעותית שהסטודנט חווה הייתה לחץ

avg_alone_percentage – כמה אחוזים מהזמן הסטודנט מבלה לבדו

avg_together_percentage – כמה אחוזים מהזמן הסטודנט מבלה בחברת אחרים

avg_working_percentage – כמה אחוזים מהזמן הסטודנט לומד

avg_relaxing_percentage – כמה אחוזים מהזמן הסטודנט לא לומד

פירוט המשתנים באקסל של הdata-set השני:

study – באיזו אוניברסיטה לומד הסטודנט, מתוך ה5 שהשתתפו בניסוי

demo_firstgen – האם הסטודנט הוא דור ראשון להשכלה (0 אם לא, 1 אם כן)

bedtime_mssd - השונות בשינה (עד כמה השינה של הסטודנט לא סדירה)

TotalSleepTime - שינה בממוצע בלילה (בדקות)

Midpoint_sleep – שעת אמצע השינה שלו, מחושבת כך – כמה דקות מפרידות בין שעה זו ל23:00 בלילה

Frac_nights_with_data – כמות לילות במהלך הניסוי בהם לא בוצעה מדידה

daytime_sleep - כמות שינה ממוצעת במהלך היום (בדקות)

score_scaled – ממוצע ציונים, מנורמל ע"פ הממוצע באותה אוניברסיטה של הסטודנט