

1. מבוא

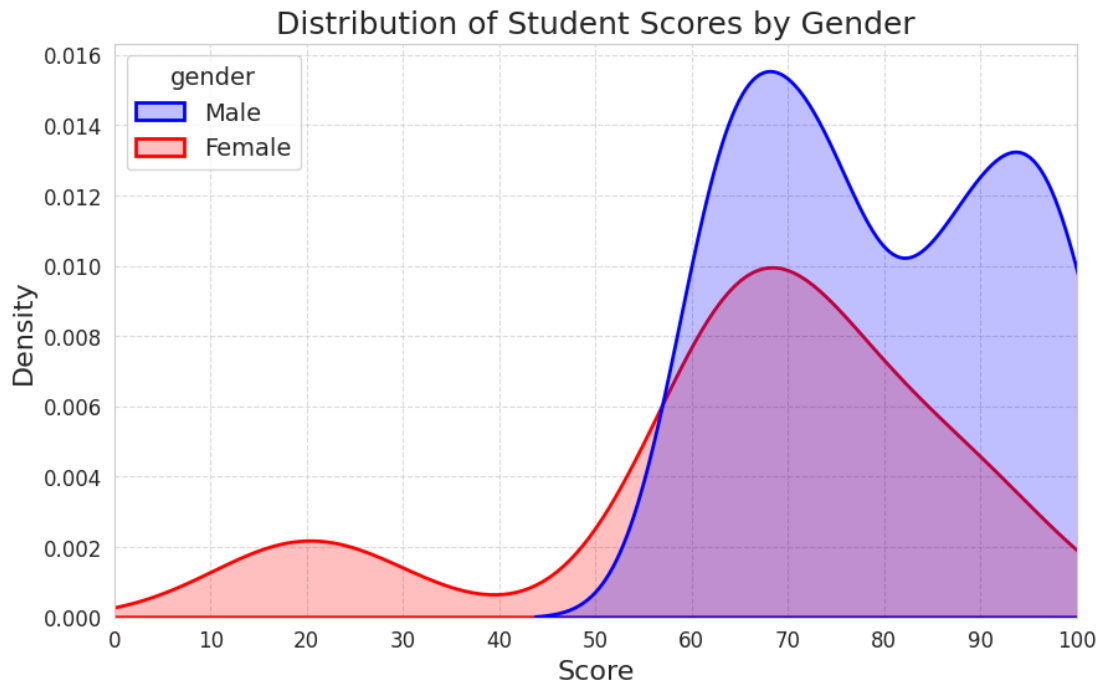
מאגר הנתונים "Higher Education Students Performance Evaluation" שנאסף בשנת 2019 מכיל מידע על 145 סטודנטים מהפקולטה להנדסה ומהפקולטה למדעים, מאוניברסיטה בקפריסין. הדאטה כולל 31 משתנים, ביניהם משתנים דמוגרפיים, הרגלי לימוד והרגלי חיים. ניתוח מאגר הנתונים עשוי לסייע להצביע על הקשר בין הצלחה בלימודים (שנמדדת בדאטה על ידי ציון בקורס) לבין משתנים דמוגרפיים וחברתיים ולהרגלי לימוד. השאלות שמעניינות אותנו קשורות להבדלים בין נשים וגברים בהצלחה בלימודים והרגלי הלימוד והעבודה ולקשר שבין השכלת ההורים להצלחתם של הסטודנטים.

בדאטה המקורי, משתנה הציון הוא משתנה אורדינלי שמקבל ערכים של בין 0-7, כאשר 0 הוגדר כ"נכשל" ואילו 7 הוגדר כציון הגבוה ביותר. על מנת לשפר את איכות ניתוח הנתונים הגדרנו מנגנון להמרת ציונים איכותיים לציונים מספריים. על פי שיטה זו, ציון מספרי בטווח 0-59 מוגדר כנכשל. יתר הציונים מחולקים לשבע קטגוריות איכותיות (מ-1 עד 7) כאשר כל קטגוריה מייצגת טווח ציונים שווה ככל הניתן. (כאשר 6 קטגוריות מכילות ערכים של 6 מספרים שלמים ואילו הקטגוריה של הציון הגבוה ביותר מכילה 5 מספרים שלמים). עבור כל סטודנט הגרלנו ציון לפי טווחי הערכים שהגדרנו על מנת לשייך ציון מספרי אמיתי (משתנה הציון עדיין אינו רציף במלואו, מפני שמקבל רק ערכים שלמים אך מייצג ערכי ציונים אמיתיים אשר ניתן לקבל). העובדה שטווחי הציונים שווים ויחסית קטנים מייצגת באופן די אמין את הציונים האמיתיים שהסטודנטים יכלו לקבל. המרה זו מאפשרת השוואה פשוטה וברורה בין ההישגים של סטודנטים, ועשויה להפוך את ניתוח הנתונים ליעיל יותר. טווחי הציונים מוצגים להלן: 0 : 0-59 ; 1 : 60-65 ; 2 : 66-71 ; 3 : 72-77 ; 4 : 78-83 ; 5 : 84-89 ; 6 : 90-95 ; 7 : 96-100.

2. התפלגות ציונים כללית

בבסיס הנתונים הקיים, בדקו 58 נשים ו-87 גברים. מתוך 145 הסטודנטים שנבדקו, 8 נכשלו וכולן נשים. הציון החציוני בקרב כלל הסטודנטים היה 73, בקרב גברים 76 ואילו בקרב נשים 68.5. התפלגות הציונים בקרב גברים ונשים מציגה מגמות שונות. אצל נשים ההתפלגות רחבה יותר מפני שיש קבוצה של ציונים נמוכים (כל 8 הנכשלות הן נשים) עם שתי פסגות (20 ו-68), ואילו אצל גברים רוב הציונים מרוכזים בטווחים גבוהים יותר עם שתי פסגות (70 ו-90). גם בטווח ציון "עובר" (60 ומעלה), ניתן לראות התפלגות שונה בין גברים לנשים.

תרשים 1: התפלגות הציונים בקרב גברים ונשים



תרשים 1. התפלגות הציונים בקרב גברים (בכחול) ונשים (באדום)

3. הרגלי למידה והבדלים בין-מגדריים

על מנת להבין טוב יותר את הבדלי הציונים בין הסטודנטים ביחס להרגלי למידה, ואת הבדלי הציונים בין גברים לנשים, נבדוק הבדלים בין הציונים החציוניים ביחס להרגלי הלמידה. מתרשים 2 ניתן לראות שיש שסטודנטים שנכחים בכיתה הם בעלי ציון חציוני גבוה יותר ($p = 0.094$) ושסטודנטים שעובדים במהלך הלימודים הם בעלי ציון חציוני נמוך יותר ($p = 0.085$). בשאר הרגלי הלמידה כגון הקשבה בכיתה, סיכום החומר והרגלי שעות הלמידה לא נמצאו הבדלים בין הקבוצות. עבור משתנים בעלי יותר מ-2 ערכים אפשריים (מקשיבים, מסכמים, מספר שעות למידה) השתמשנו במבחן Kruskal-Wallis אשר בודק האם יש לפחות קבוצה אחת ששונה באופן מובהק משאר הקבוצות.

עבור המשתנים בהם היו הבדלים בין הקבוצות, נוכחות בכיתה ועבודה במהלך הלימודים, וגם עבור המשתנה מספר שעות לימוד שבועיות (אשר לו מצאנו שאין השפעה על הציון אבל הגיוני לשער שדווקא כן יש לו), בדקנו את הציונים החציוניים בהפרדה מגדרית.

נוכחות בשיעורים

מתרשים 3 עולה כי גברים שנכחו בשיעורים קיבלו ציון חציוני גבוה מאשר נשים שנכחו בשיעורים. אפקט זה לא משתחזר בתנאי "ללא נוכחות בשיעור", כאשר הציון החציוני של גברים ונשים דומה יותר עם חפיפה גדולה בטווחי הציונים. בנוסף, ניתן לראות שההשפעה של נוכחות בשיעורים מקורה בעיקר בגברים, כאשר כמעט ואין הבדל בין הציון החציוני של נשים בשתי הקבוצות אך כן יש הבדל בקרב גברים שנכחו בשיעור (ציון גבוה יותר).

עבודה במהלך הלימודים

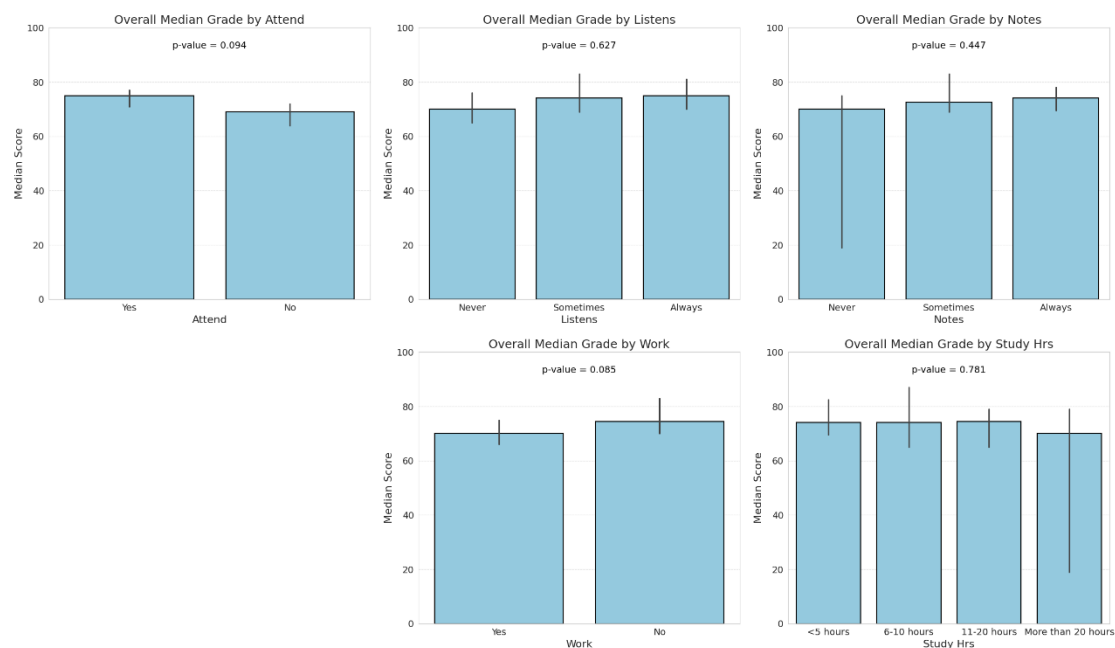
מתרשים 4 עולה כי בשתי קבוצות העבודה (כן עובדים, לא עובדים) גברים קיבלו ציונים חציוניים טובים יותר מאשר הנשים. בנוסף, נראה כי בקרב הנשים, עבודה במהלך הלימודים לא משפיעה על ההישגים בלימודים ואילו בקרב הגברים עבודה בלימודים קשורה בירידה קלה בהישגים.

מספר שעות לימוד שבועיות

בקרב גברים, הציונים הגבוהים יותר מתקבלים דווקא בקבוצה של מספר שעות הלמידה השבועיות הנמוך ביותר (5 שעות ומטה) כאשר החציון בקבוצה זאת גבוה במיוחד. בנוסף, בקרב הגברים, לא נראה שיש הבדל ככל שלומדים יותר מ-5 שעות. אצל נשים נראה שלמידה של בין 11-20 שעות מפיקה את הציונים הגבוהים ביותר. בנוסף, הציונים של הנשים נמוכים משל הגברים בקבוצות "5 שעות ומטה" ו-"6-10", ואילו מעט גבוהים יותר כאשר מספר שעות הלימוד השבועיות הוא בין 11-20 שעות. קבוצת "20 שעות ומעלה" מונה 4 סטודנטים בסה"כ, מתוכם 3 נשים וגבר אחד בלבד כך שהשוואה זאת אינה מלמדת אותנו הרבה. מתרשים 5 אנו למדים שאין קשר ברור בין מספר שעות הלמידה השבועיות לבין הציון הסופי, במיוחד אצל זכרים, כאשר הציונים הגבוהים ביותר מתקבלים דווקא בקבוצות של מעט שעות למידה.

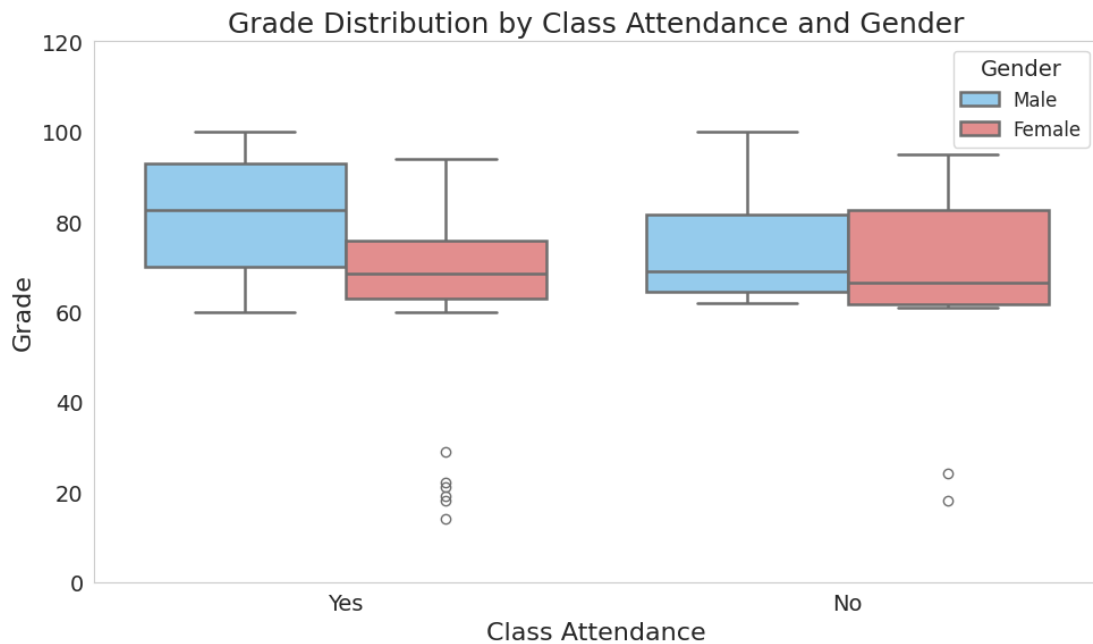
תרשים 2: ציונים חציוניים לפי נוכחות, הקשבה, סיכום, עבודה בזמן לימודים ומספר שעות

לימוד שבועיות



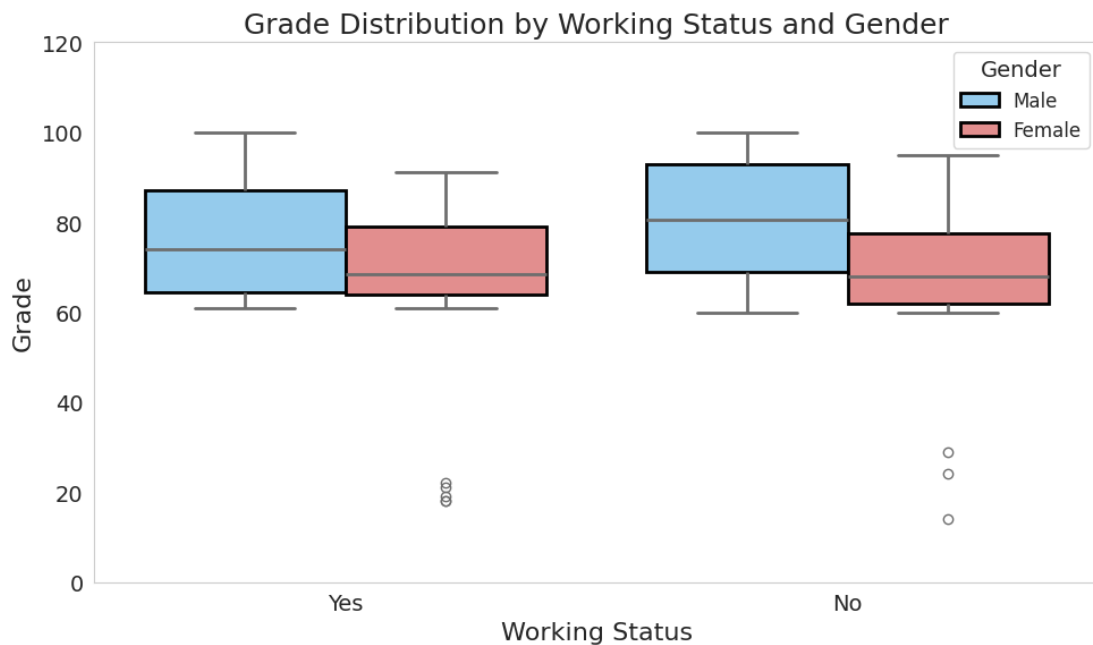
תרשים 2. הציונים החציוניים לפי הרגלי למידה שונים ושגיאות התקן.

תרשים 3: התפלגות הציונים לפי נוכחות בכיתה ומגדר



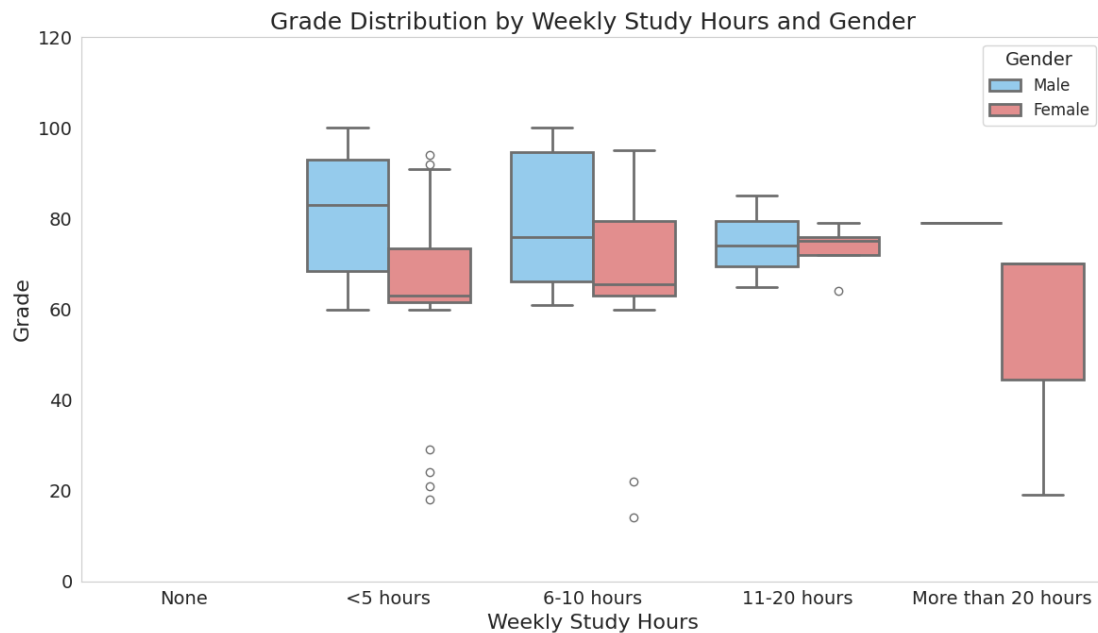
תרשים 3. התפלגות הציונים בקרב גברים (בכחול) ונשים (באדום) לפי נוכחות בכיתה (נוכח/ת – Yes, לא נוכח/ת – No)

תרשים 4: התפלגות הציונים לפי סטטוס עבודה במהלך הסמסטר ומגדר



תרשים 4. התפלגות הציונים בקרב גברים (בכחול) ונשים (באדום) לפי סטטוס עבודה במהלך הסמסטר (עובד/ת – Yes, לא עובד/ת – No)

תרשים 5: התפלגות הציונים לפי נוכחות בכיתה ומגדר



תרשים 5. התפלגות הציונים בקרב גברים (בכחול) ונשים (באדום) מספר שעות לימוד שבועיות

4. השכלה של ההורים

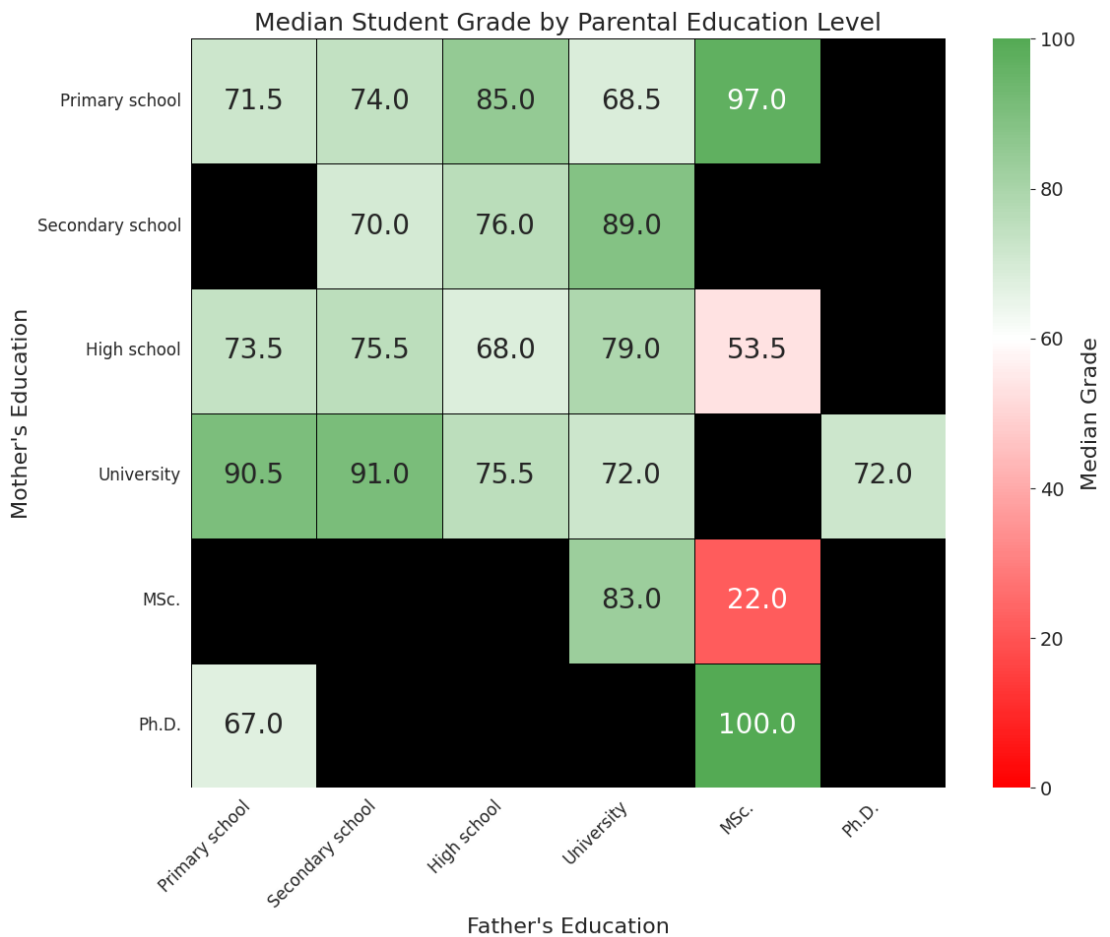
אחד הדברים המעניינים הוא הקשר בין השכלת הסטודנטים לבין השכלת ההורים. בדקנו האם יש קשר בין ההישגים בלימודים לבין השכלת ההורים. בנתונים יש את פירוט השכלת ההורה מבית ספר יסודי ועד לתואר דוקטור. לפי תרשים 6 לא ניתן להגיד שבקרב הסטודנטים והסטודנטיות שנבדקו ישנו קשר בין רמת ההשכלה של ההורים לבין ההישגים בלימודים.

תרשים 6 מציג את הציון החציוני של סטודנטים לפי רמת ההשכלה של האם והאב. מהתרשים עולה שאין דפוס עקבי שבו עלייה ברמת ההשכלה של אחד ההורים או שניהם מובילה בהכרח לציונים גבוהים יותר אצל הסטודנטים. ניתן לראות דוגמאות לציונים גבוהים גם כאשר להורים השכלה נמוכה, ולהפך – יש מקרים שבהם להורים השכלה גבוהה במיוחד, אך הציונים החציוניים של ילדיהם דווקא נמוכים יחסית. בנוסף, ישנם שילובים מסוימים של השכלת הורים שבהם הציונים החציוניים גבוהים במיוחד, אך אין בכך כדי להצביע על מגמה ברורה.

תרשים 7, המציג את התפלגות הציונים לפי קבוצות השכלה של ההורים (שני הורים בעלי השכלה גבוהה, שני הורים בעלי השכלה נמוכה, או שילובים ביניהם), מחזק את הממצא הזה. הפערים בין הקבוצות אינם משמעותיים, ופיזור הציונים רחב בכל אחת מהקבוצות. כלומר, גם כאשר שני ההורים משכילים, יש שונות רבה בציוני הסטודנטים, ואין עדות לכך שלהשכלת ההורים יש השפעה ברורה ועקבית על הישגי הילדים.

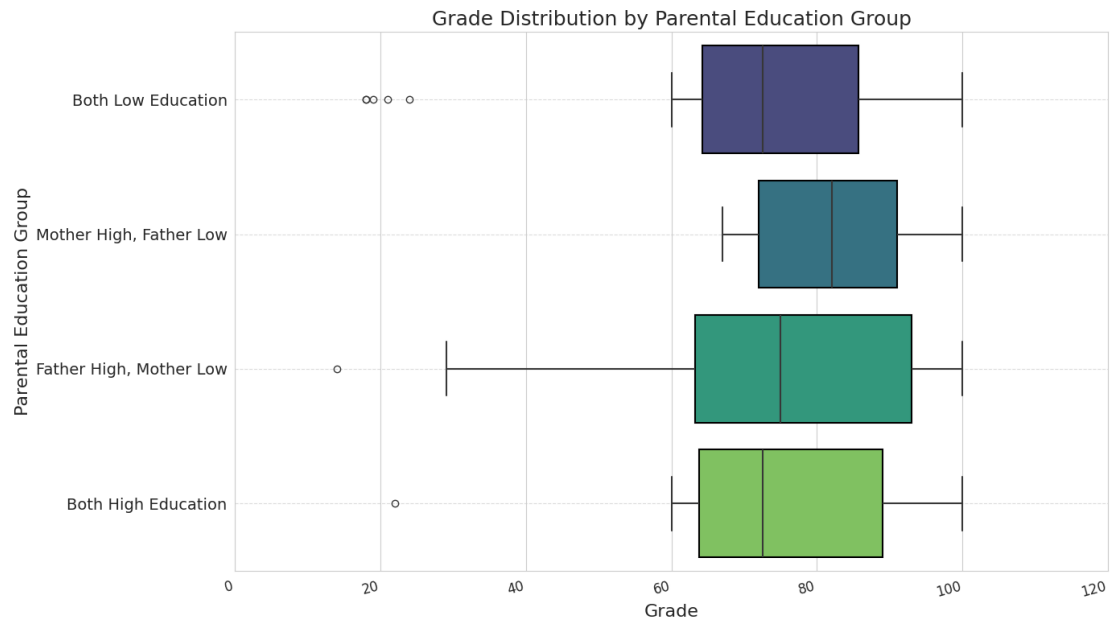
שני התרשימים יחד מצביעים על כך שבקרב הסטודנטים שנבדקו, לרמת השכלת ההורים קשר להישגים אקדמיים.

תרשים 6: ציון חציוני של סטודנטים לפי רמת השכלה של האם והאב



תרשים 6. מפת חום המציגה את הציון החציוני של סטודנטים בהתאם לרמת ההשכלה של האם (שורות) והאב (עמודות). ערך הביניים (בצבע לבן) הוא ציור עובר (60), כאשר ציונים גבוהים יותר נצבעים בצבע ירוק ואילו ציונים נמוכים יותר נצבעים בצבע אדום.

תרשים 7: התפלגות הציונים לפי קבוצות השכלה של ההורים



תרשים 7. התפלגות הציונים בקרב קבוצות ההשכלה של ההורים.

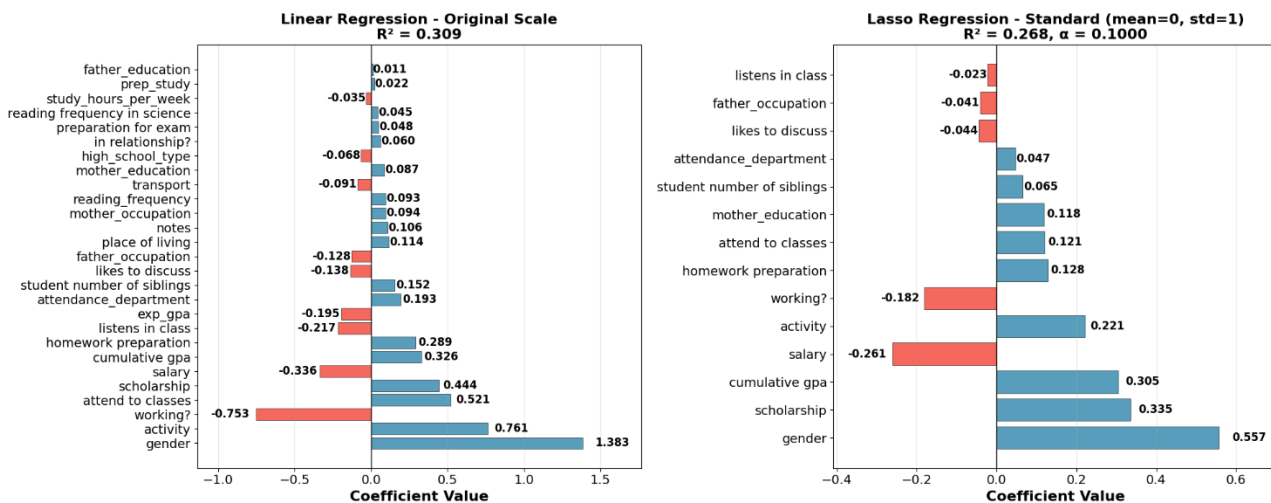
5. אמידת פרמטרים

בחלקים הקודמים בעבודה שאלנו האם יש מאפיינים ספציפיים מסוימים אשר יכולים להיות קשורים להצלחה בלימודים. עמדנו על ההבדלים בין מגדרים, הרגלי למידה ורקע השכלתי משפחתי. בסיס הנתונים כולל משתנים נוספים אשר יכולים להיות קשורים להצלחה בלימודים ולכן בחלק זה נבצע אמידת פרמטרים בעזרת רגרסיה לינארית. כאמור, הדאטה כולל 31 משתנים (מתוכם, לאחר סינון נשארו 27) ועל מנת לנסות להימנע מ-overfitting, ביצענו גם רגרסיית Lasso. ישנם כמה משתנים אשר משתנים אשר לא נכללו ברגרסיית Lasso לעומת הרגרסיה הלינארית הרגילה, למשל, מספר שעות הלמידה השבועיות (אשר ראינו כבר היעדר קשר), השכלת האב, סוג בית הספר התיכון בו למדו. כמו כן, אחוז השונות המוסבר ברגרסיה הלינארית היה 0.309. ברגרסיית ה-Lasso (עם פרמטר אלפא שווה ל-0.1), נשארו 14 משתנים רלוונטיים, כאשר מחצית מתוכם מובהקים סטטיסטית ברמת מובהקות של $p < 0.05$. המשתנה עם רמת השפעה הגדולה ביותר על הציון הסופי, וברמת המובהקות הגבוהה ביותר הוא משתנה המגדר. כמובן שאין לחשוב שגברים מצליחים יותר בלימודים מאשר נשים, אלא זוהי תוצאה של המדגם הספציפי הזה.

שתי תוצאות מעניינות עולות מהמשתנים: 1) האם הסטודנטית עובדת?; גובה המלגה; תוצאות האמידה מצביעות על השפעה שלילית על הציון הסופי של עבודה במהלך הלימודים, ובמקביל על השפעה חיובית של גובה המלגה על הציון הסופי. ישנה מחלוקת לגבי התרומה של עבודה במהלך הלימודים לסטודנטים. מחד, סטודנטים שעובדים משקיעים פחות זמן בלימודים וכך הציון שלהם יכול לרדת, ומאידך, עבודה עשויה לתרום הרבה באספקטים אחרים שלא קשורים בהכרח לציון אבל כן משפיעים על הסטודנטים. במדגם אותו חקרנו מצאנו

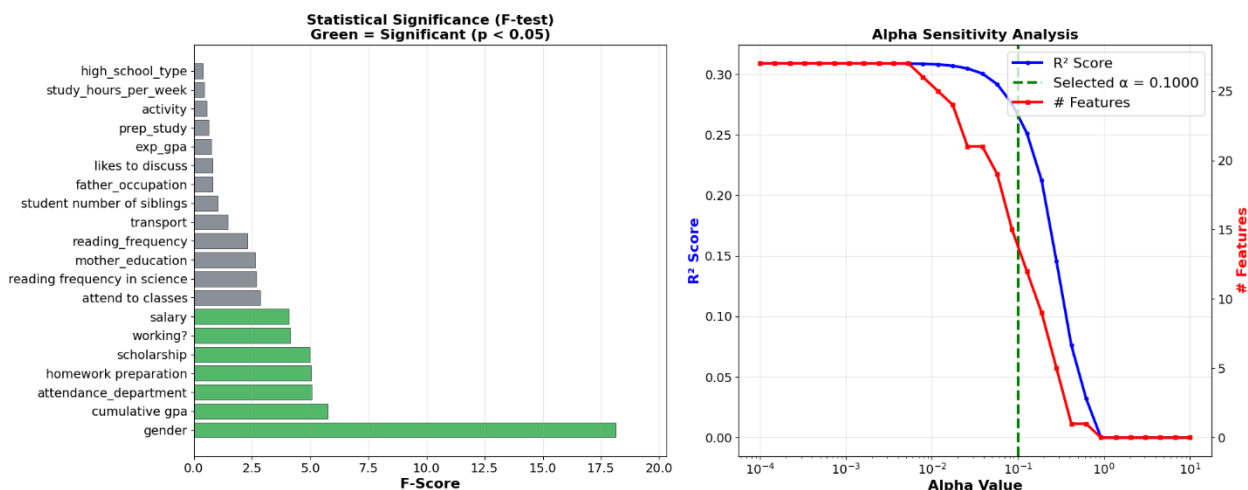
שלעבודה במהלך הלימודים יש השפעה שלילית על הציון הסופי בקורס. באופן לא מפתיע, גם משתנה השכר (המתואם באופן חיובי עם עבודה במהלך הלימודים) מתואם שלילית עם הציון הסופי ומשפיע עליו לרעה. אחוז השונות המוסברת ברגרסיית Lasso נמוך יותר ועומד על 0.268. חשוב לומר, שהמשתנים שיצאו מובהקים סטטיסטית (ברמת מובהקות מקובלת של 5%), הם המנבאים הטובים ביותר לציון הסופי רק על פי הנתונים האלו, ויהיה קשה להכליל ולהגיד שאלו הם המנבאים הטובים ביותר באופן כללי.

תרשים 8: תוצאות הרגרסיות השונות (לינארית ו-lasso)



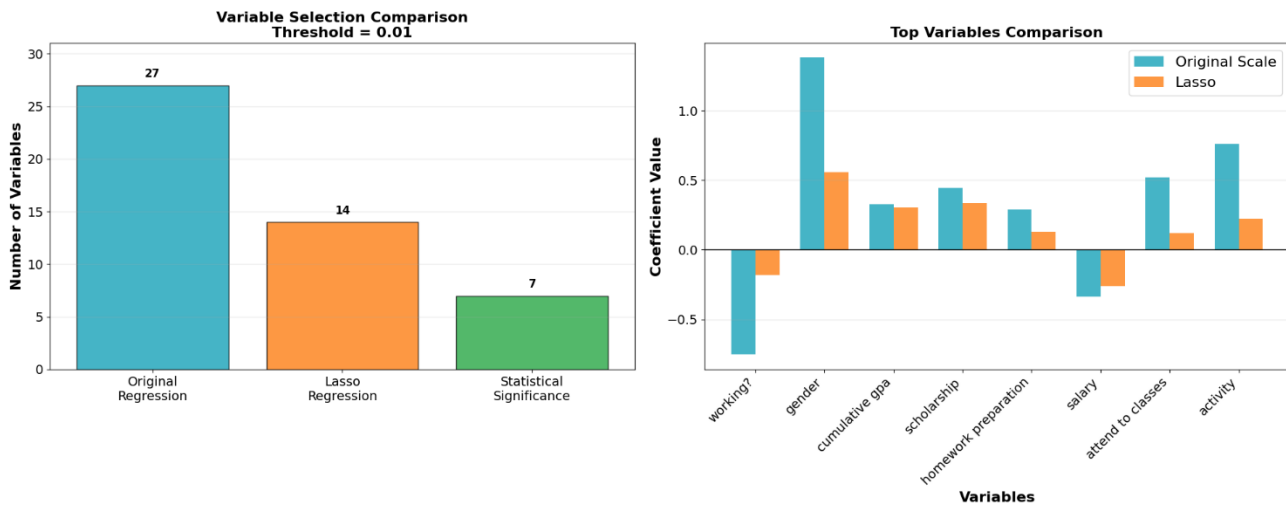
תרשים 8. התרשימים מציגים אמידה של הציון הסופי אל מול מגוון משתנים מסבירים. התרשים השמאלי מציג רגרסיה לינארית ואילו התרשים הימני מציג רגרסיית lasso אשר מספר המשתנים המסבירים בה יורד ל-14.

תרשים 9: מובהקות התוצאות ורגישות רגרסיית lasso לגודל הפרמטר אלפא



תרשים 9. התרשימים מציגים את מובהקות הפרמטרים לפי מבחן F (להתאמת פרמטרים) ואת רגישות כמות המשתנים שנכללים ברגרסיית lasso ואת טיב ההתאמה לפי גודל הפרמטר אלפא.

תרשים 10: השוואה בין רגרסיה לינארית לרגרסיית lasso



תרשים 10. ברגרסיה הלינארית כללנו 27 משתנים מסבירים, אך ברגרסיית lasso כמות המשתנים המסבירים ירדה ל-14 כאשר מותכם 7 מובהקים סטטיסטית ברמת מובהקות של 5%. בתרשים הימני ניתן לראות השוואה של ערכי המקדמים (המנורמלים) בין הרגרסיה הלינארית לרגרסיית lasso.

6. סיכום

בפרויקט זה ניסינו לבדוק אילו גורמים משפיעים על הצלחה בלימודים, אשר נמדדה, בסיס הנתונים הזה, על ידי ציון סופי במבחן. מצאנו תוצאות די מפתיעות, ולדעתנו, לא מייצגות את האוכלוסייה הכללית. כלומר התוצאות שמצאנו מייצגות לדעתנו רק את המדגם אותו בדקנו. מצאנו את כל שהמשתנים שמנבאים בצורה הטובה ביותר את הציון הסופי בקורס כהגיוניים ביותר, אך עם זאת, מצאנו גם שישנם גם משתנים שנמצאו שלא משפיעים על הציון הסופי אך זה פחות עושה היגיון עם המציאות, ולכן לדעתנו התוצאות שמצאנו מייצגות רק את המדגם הזה.