פרויקט גמר- R למתחילים | תהילה מן 315605857

קישור להצגה מלאה של הניתוח: https://github.com/Tehilla-TAU/CS_Grad_Project

****

שלב א' - הגדרת שאלת המחקר

בחירת דאטאסט:

בחרתי לנתח את הדאטאסט ,"**Education & Career Success**" שנלקח ממאגר הנתונים Kaggle.

הדאטאסט שנבחר עוסק בקשר בין השכלה להצלחה בקריירה, תוך התמקדות בגורמים שעשויים להשפיע על השכר. הדאטאסט כולל מגוון משתנים אקדמיים ומקצועיים, כגון דירוג האוניברסיטה, ממוצע ציונים, מספר ההתמחויות, כישורי נטוורקינג ועוד. הדאטה כולל נתונים של כ-5,000 בוגרים מתחומים שונים.

במחקר זה בחרתי להתמקד אך ורק בבוגרי מדעי המחשב. מניסיוני בתחום התעסוקה, ההשפעה של ההשכלה על הקריירה שונה באופן מהותי בין תחומי עיסוק שונים. ישנם מקצועות בהם תואר אקדמי הוא הכרחי על מנת לקבל רישיון לעסוק בתחום, כמו רפואה או עריכת דין. לעומת זאת, בתחום מדעי המחשב אין חובה כזו – וניתן להיכנס לתעשייה גם ללא תואר, דרך קורסים מקצועיים והכשרות שונות. בנוסף, ישנם מוסדות לימוד רבים המציעים מסלולי מדעי המחשב ברמות אתגר שונות, ולכן עולה השאלה מהם הגורמים המשמעותיים ביותר שמשפיעים בפועל על השכר ואפשרויות התעסוקה של בוגרי התחום.

לאחר סינון סטודנטים שלמדו **מדעי המחשב בלבד** ,נשארו **670 תצפיות** שנותחו במחקר.

הסיבה האישית לבחירה בדאטאסט זה נעוצה בכך שכיום אני עוסקת בתפקיד הקשור בשילוב חרדים בשוק העבודה, ומתעניינת בהבנת ההשפעה של הרקע הלימודי על הצלחה תעסוקתית. כמו כן, יש לי עניין בזיהוי הגורמים המשמעותיים שיכולים לסייע לסטודנטים – ובפרט לאוכלוסיות מוחלשות – למקסם את סיכוייהם להשתלב במשרות איכותיות. מחקר זה עשוי לספק תובנות חשובות לגבי תהליכי השכלה והשלכותיהם על הקריירה.

הצגה מקדימה של נתונים(מצורפים הגרפים , בסוף עמ' 2):

לפני ביצוע הניתוחים הסטטיסטיים, בוצעה חקירה ראשונית של הנתונים בעזרת החבילות **ggplot2** ו ,**ggdist**- תוך הצגת התפלגות המשתנים והקשרים האפשריים ביניהם (הגרפים מצורפים בסוף עמ' 2).

שכר התחלתי : השכר החציוני בקרב בוגרי מדעי המחשב הוא כ-$50,000, עם התפלגות סימטרית יחסית. קיימת שונות גבוהה בין הסטודנטים, עם טווח שכר רחב (בין $30,000 ל-$90,000).

מספר ההתמחויות: קיימת שונות משמעותית בניסיון המעשי בהתמחויות. חלק מהסטודנטים ללא התמחויות כלל, בעוד אחרים צברו 3–4 התמחויות. נתון זה מעלה עניין באפשרות שהניסיון המעשי עשוי להיות גורם שמשפיע על השכר.

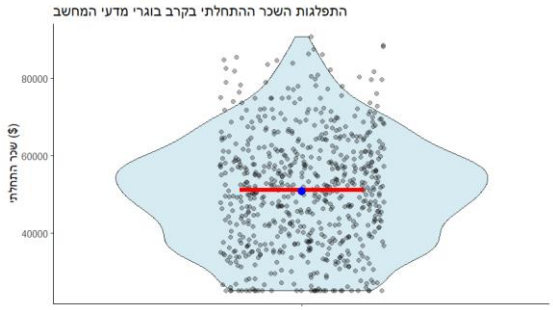
דירוג האוניברסיטה : לא נראה קשר ברור בין דירוג האוניברסיטה לשכר ההתחלתי. קיימת שונות גבוהה בכל רמות הדירוג, מה שמעורר עניין לבחון משתנים אחרים, מעבר לדירוג המוסד.

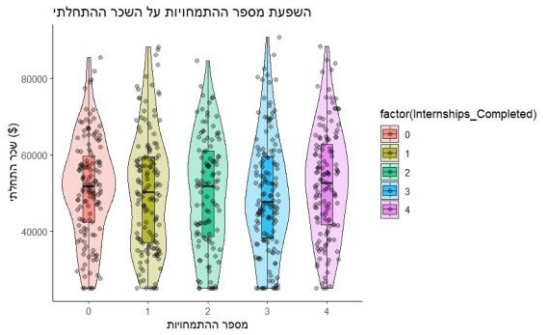
כישורי נטוורקינג : נראית כי קיימת שונות בין הסטודנטים מבחינת כישורי הנטוורקינג שלהם. על אף שבשלב זה טרם נבדק הקשר לשכר, ההשערה היא שכישורים אלו עשויים להיות בעלי משמעות רבה בעולם העבודה, ובפרט בתחום מדעי המחשב שבו כישורים חברתיים יכולים לפתוח הזדמנויות תעסוקתיות.

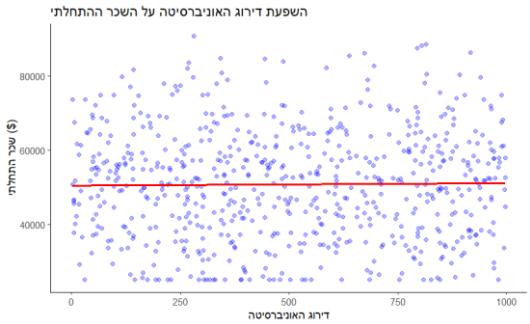
ניסוח שאלת המחקר:

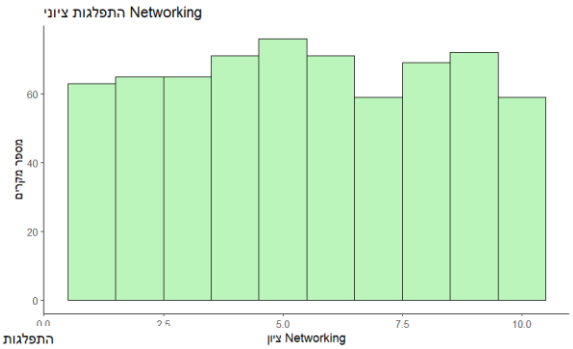
To what extent do the number of internships and networking skills predict the starting salary of computer science graduates? Additionally, can these factors accurately classify graduates as earning above or below the median salary?

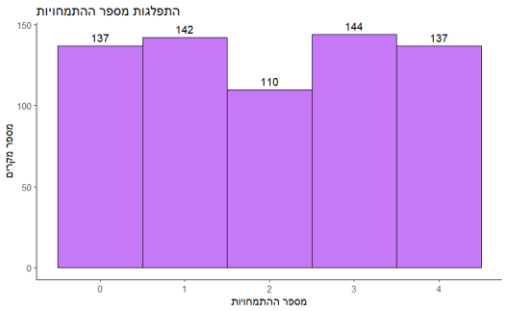
<u>הצגת נתונים מקדימה- גרפים</u>

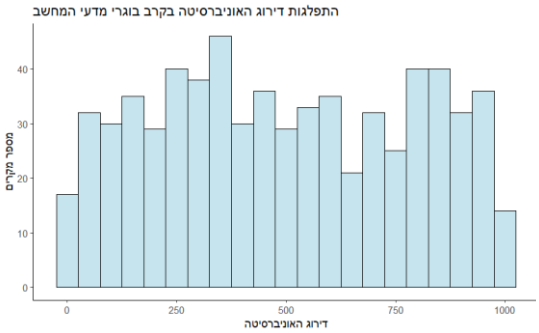


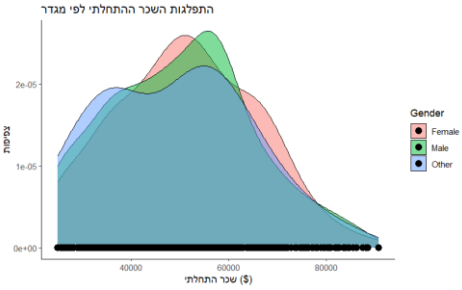












שלב ב' - עיבוד מקדים של הנתונים

הגדרת המשתנים לניתוח:

משתנה תלוי:

Starting_Salary משתנה רציף המייצג את גובה השכר ההתחלתי של הבוגר (נמדד בדולרים). ישמש כמשתנה המנובא במודל רגרסיה לינארית.

משתנה תלוי לרגרסיה לוגיסטית:

High_Salary משתנה בינארי הקובע אם הבוגר מרוויח מעל או מתחת לחציון השכר .

0= שכר נמוך או שווה לחציון, 1 = שכר גבוה מהחציון. ישמש כמשתנה המנובא במודל רגרסיה לוגיסטית.

משתנים בלתי תלויים (מנבאים):

Internships_Completed משתנה כמותי המייצג את מספר ההתמחויות שביצע הבוגר במהלך לימודיו.

Networking_Score משתנה כמותי המייצג את דירוג כישורי הנטוורקינג של הבוגר (ציון מדורג).

*יצירת משתנה עזר לצורך המחשה גרפית בלבד:

Networking_Level משתנה קטגורי שנוצר מתוך Networking_Score ומחלק אותו לשלוש רמות: Low - שלישון תחתון, Medium- שלישון אמצעי, High- שלישון עליון.

המשתנה מאפשר לבחון האם קשרי נטוורקינג בדרגות שונות משפיעים על השכר או על סיכוי להשתכרות גבוהה.

עיבוד הנתונים באמצעות: dplyr

הנתונים עברו ניקוי וסינון לקראת הניתוח:

- סינון לסטודנטים שלמדו מדעי המחשב בלבד: filter
- השלמת ערכים חסרים באמצעות חציון: mutate, across, median
- הסרת חריגים על בסיס IQR (quantile()) בשילוב filter, across כדי למנוע עיוותים בניתוח הרגרסיה
- שמירה רק על המשתנים הרלוונטיים למחקר: select

פונקציה מותאמת אישית

לצורך בניית משתנה בינארי המסווג את השכר לקטגוריות גבוה/נמוך, נכתבה פונקציה ייעודית:

הפונקציה **מחשבת את חציון השכר** ומסמנת כל תצפית בהתאם:

1= שכר מעל החציון, 0= שכר נמוך או שווה לחציון

ניקוי עמודות עם ערכים חסרים (בונוס)

במסגרת הבונוס, נעשה שימוש בחבילת janitor לניקוי אוטומטי של הנתונים:

- פונקציה clean_missing_data() מסירה עמודות שבהן יותר מ-20% ערכים חסרים. היא מאפשרת צמצום עמודות לא שימושיות ושיפור איכות הנתונים לפני עיבוד נוסף. הקוד מופעל בתחילת התהליך, לפני טיפול בערכים חסרים והסרת חריגים.

שלב ג' - ניתוח הנתונים

רגרסיה לינארית מרובה: ניבוי השכר ההתחלתי (Starting_Salary)

נבדק האם מספר ההתמחויות וכישורי הנטוורקינג מנבאים את השכר ההתחלתי.
ממצאים עיקריים:

מספר ההתמחויות: $\beta = 157.7$, $p = 0.689$ לא מובהק.

כל התמחות נוספת מעלה את השכר ההתחלתי ב-157.7$ בממוצע, אך ללא מובהקות סטטיסטית.

כישורי נטוורקינג: $\beta = 192.4$, $p = 0.339$ לא מובהק.

כל נקודה נוספת במדד הנטוורקינג מעלה את השכר ההתחלתי ב-192.4$ בממוצע, אך ללא מובהקות סטטיסטית.

מדדי התאמה למודל:

Residual SE = 14,650 מעיד על שונות גבוהה שאינה מוסברת ע"י המודל.

R-squared = 0.0016 המודל מסביר רק 0.16% מהשונות בשכר ההתחלתי

אין קשר מובהק בין מספר ההתמחויות או כישורי הנטוורקינג לבין השכר ההתחלתי. המודל אינו מסביר את השונות בשכר

רגרסיה לוגיסטית: ניבוי הסיכוי לשכר מעל החציון (High_Salary)

הניתוח בדק עד כמה ניתן לחזות האם בוגר ירוויח מעל או מתחת לחציון השכר, בהתבסס על מספר ההתמחויות וכישורי הנטוורקינג.
ממצאים עיקריים:

מספר ההתמחויות: $\beta = 0.0079$, $p = 0.883$ (לא מובהק)

כל התמחות נוספת מגדילה את log-odds לשכר גבוה ב 0.0079 בלבד .

$Exp(\beta) = e^{0.0079} \approx 1.008$ כלומר עלייה של 0.8% בסיכוי לשכר גבוה ללא מובהקות סטטיסטית.

כישורי נטוורקינג: $\beta = 0.0357$, $p = 0.193$ (לא מובהק)

כל נקודה נוספת במדד הנטוורקינג מגדילה את log-odds לשכר גבוה ב 0.0357

$Exp(\beta) = e^{0.0357} \approx 1.036$ כלומר עלייה של 3.6% בסיכוי לשכר גבוה – ללא מובהקות סטטיסטית.

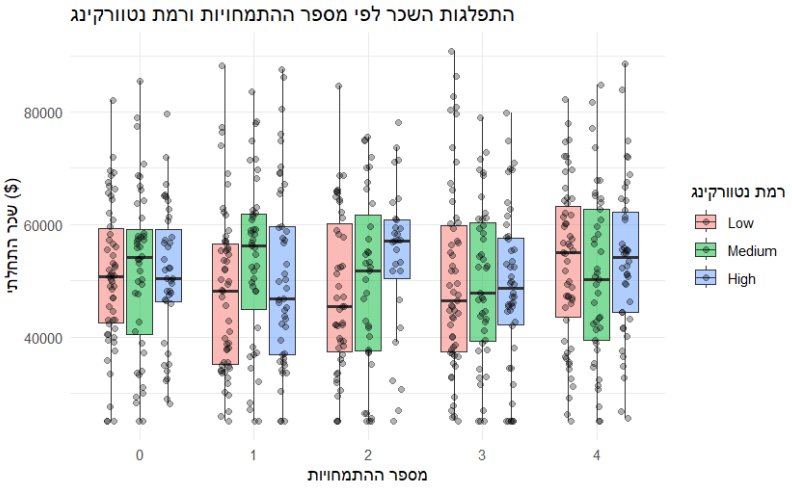
$Exp(\beta)$ קרוב ל-1, מה שמעיד על כך שאין השפעה משמעותית של מספר ההתמחויות או כישורי הנטוורקינג על רמות השכר.

מדד התאמת המודל: (AIC) = 933.1

ערך גבוה יחסית, מה שמעיד על התאמה נמוכה של המודל לנתונים.

שטח מתחת לעקומת ROC: (AUC) = 0.53 מה שמעיד כי המודל אינו משפר משמעותית את הניבוי בהשוואה לניחוש אקראי (0.50) .

גרפים לתיאור כל אחד מהאפקטים שהתקבלו



אפקט מספר ההתמחויות וכישורי הנטוורקינג על השכר (עבור הרגרסיה הלינארית)

גרף מציג את התפלגות השכר ההתחלתי לפי מספר ההתמחויות ורמות כישורי הנטוורקינג/

-פיזור רחב מאוד של השכר בתוך כל קבוצה ,ללא מגמה ברורה של עלייה בשכר עם מספר ההתמחויות.

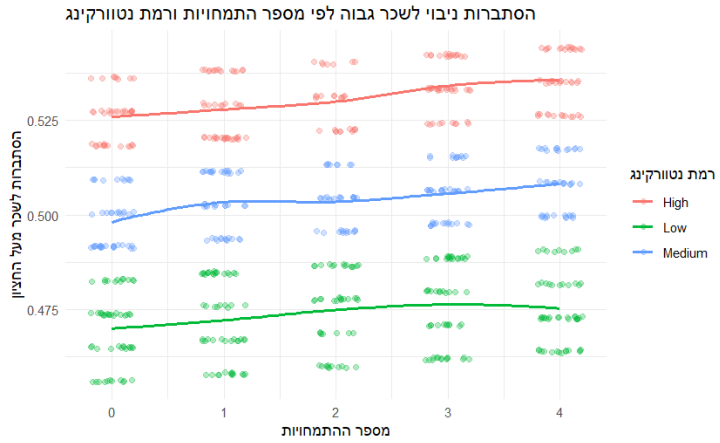
-החציון דומה בין הקבוצות ,מה שמעיד על כך שלא ניתן לראות דפוס עקבי של עלייה בשכר.

-חפיפה גבוהה בין קבוצות הנטוורקינג –השכר של בעלי כישורי נטוורקינג נמוכים חופף לשכר של בעלי כישורים גבוהים, ללא הפרדה ברורה.

-נקודות חריגות רבות מצביעות על שונות גבוהה שאינה מוסברת על ידי מספר ההתמחויות או רמת הנטוורקינג.

מסקנה:

הנתונים מצביעים על היעדר קשר ברור בין מספר ההתמחויות וכישורי הנטוורקינג לבין השכר ההתחלתי.



הסתברות חזויה לשכר גבוה לפי מספר ההתמחויות ורמת הנטוורקינג (עבור הרגרסיה הלוגיסטית)

הגרף מציג את ההסתברות החזויה לשכר גבוה (מעל החציון), בהשוואה למספר ההתמחויות ולרמות כישורי הנטוורקינג (Low, Medium, High).

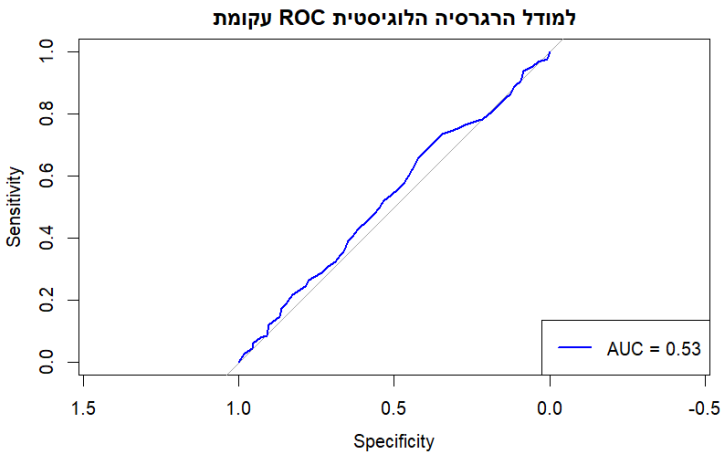
- המודל מראה שינוי קל בהסתברות לשכר גבוה עם עלייה במספר ההתמחויות ,אך מדובר במגמה חלשה מאוד.

-הפערים בין רמות הנטוורקינג אינם חד משמעיים ,ואין עדות להשפעה משמעותית של כישורים אלו על הסיכוי להרוויח מעל החציון.

-ההסתברויות נעות בטווח נמוך מאוד ,(0.47–0.53)ללא שונות משמעותית.

מסקנה:

הניתוח מראה כי מספר ההתמחויות וכישורי הנטוורקינג אינם מספקים ניבוי משמעותי של הסיכוי להשתכרות גבוהה.



גרף ROC להערכת איכות המודל הלוגיסטי

גרף ה ROC-מציג את יכולת המודל הלוגיסטי להבחין בין שכר גבוה (מעל החציון) לשכר נמוך או שווה לחציון.

AUC = 0.53מה שמעיד על כך שהמודל אינו משפר משמעותית את הדיוק בניבוי השכר לעומת ניחוש אקראי.(0.50)

הקו הכחול קרוב מאוד לקו האלכסון (ניחוש אקראי), מה שמראה שהמודל אינו מצליח להפריד היטב בין שתי הקבוצות.

מסקנה:

-המודל הלוגיסטי אינו מספק ניבוי טוב לשכר גבוה בהתבסס על מספר ההתמחויות או כישורי הנטוורקינג.

-אין עדות לכך שמשתנים אלו תורמים להבנת ההבדלים ברמות השכר.