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", אשר נלקח ממאגר הנתונים Kaggle.

הדאטאסט שנבחר עוסק בקשר בין השכלה לבין הצלחה 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 5000 בוגרים בתחומים 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, במדעי 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, מה שמעלה את השאלה: מהם הגורמים שבאמת ישפיעו על אפשרויות השכר והתעסוקה של בוגרי התחום.

הסיבה האישית לבחירה בדאטאסט זה נעוצה בכך שכיום אני עוסקת בתפקיד שקשור בשילוב חרדים בשוק העבודהומתעניינת בהבנת ההשפעה של רקע לימודי על הצלחה תעסוקתית, כמו גם 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 **ggplot2** ו-**ggdist**, תוך שימוש במגוון גרפים להמחשת התפלגות המשתנים וקשרים אפשריים ביניהם.

<u>תובנות ראשוניות המבוססות על הצגת הנתונים</u>

השכר ההתחלתי בקרב בוגרי מדעי המחשב: השכר החציוני של בוגרי מדעי המחשב עומד על כ-50,000$. השכר הממוצע קרוב מאוד לחציון, מה שמעיד על התפלגות סימטרית יחסית. טווח השכר רחב מאוד: ניתן לראות משרות התחלתיות ב-30,000$, לצד משרות עם שכר גבוה מ 90,000$ יש שונות גבוהה, כלומר קיימים פערים משמעותיים בין הבוגרים.

רמות הניסיון של הסטודנטים בהתמחויות שונות משמעותית –חלקם ללא ניסיון מעשי כלל, בעוד אחרים צברו 3-4 התמחויות.

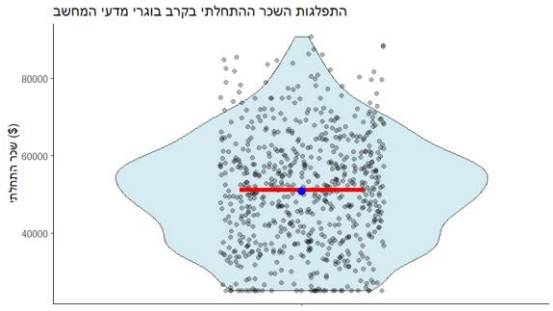
לא נראה שיש קשר חזק בין דירוג האוניברסיטה לשכר ההתחלתי .עם זאת, נראה כי קיימת שונות גבוהה מאוד בשכר בקרב כל הדרגים ,מה שעשוי להעיד כי גורמים נוספים (כגון ניסיון מעשי, רשת קשרים, או מיומנויות רכות) משפיעים יותר על רמות השכר.

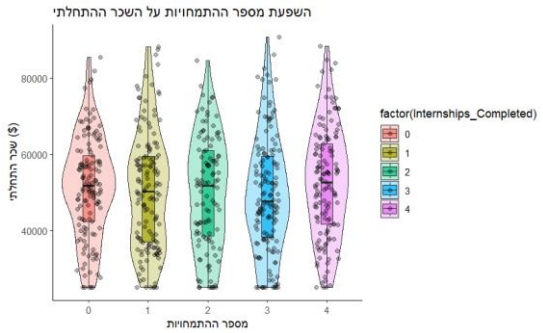
נראה כי מספר ההתמחויות עשוי להיות גורם משמעותי בשכר ההתחלתי ,שכן נראה שיש מגמה מסוימת של שכר גבוה יותר עם יותר התמחויות .עם זאת ,השונות גדולה מאוד בתוך כל קבוצה , ולכן ייתכן כי משתנים נוספים משפיעים על הקשר.

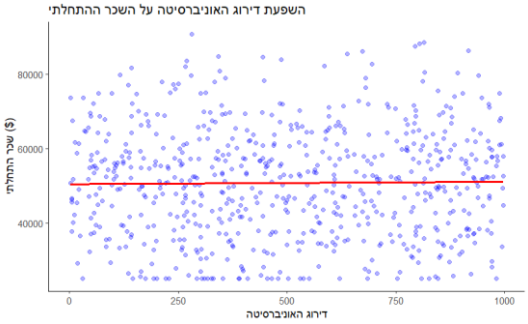
ניסוח שאלת המחקר:

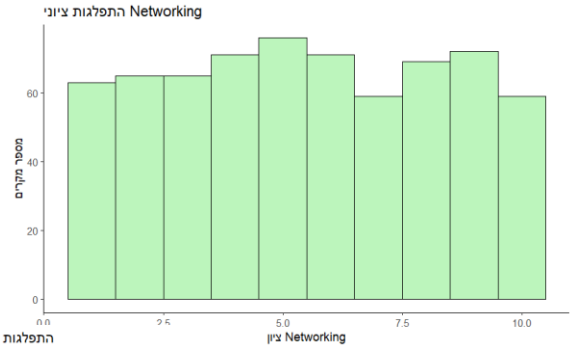
To what extent do the number of internships and networking skills predict the starting salary of computer science graduates? Additionally, can these factors accurately classify graduates as earning above or below the median salary?

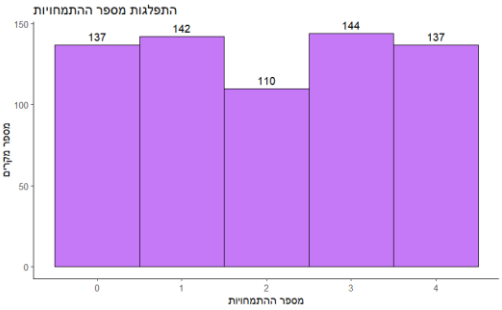
הצגת נתונים מקדימה- גרפים

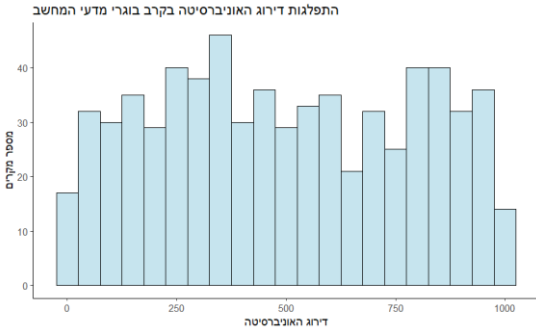


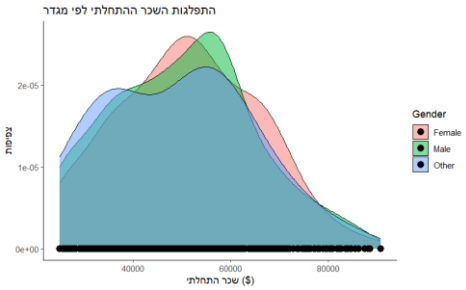












שלב ב' - עיבוד מקדים של הנתונים

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).

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

יצירת משתנה בינארי לשכר גבוה

נכתבה פונקציה ליצירת המשתנה: High_Salary

- הפונקציה מחשבת את חציון השכר ומסמנת כל תצפית כשייכת לקבוצה:
  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 ב-157.7$ בממוצע, אך לא באופן מובהק סטטיסטית

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

כל נקודה נוספת במדד הנטוורקינג מעלה את השכר ב-192.4$ בממוצע, אך לא באופן מובהק סטטיסטית.

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 האם מספר ההתמחויות וכישורי הנטוורקינג מנבאים את הסיכוי להרוויח שכר מעל החציון. 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.

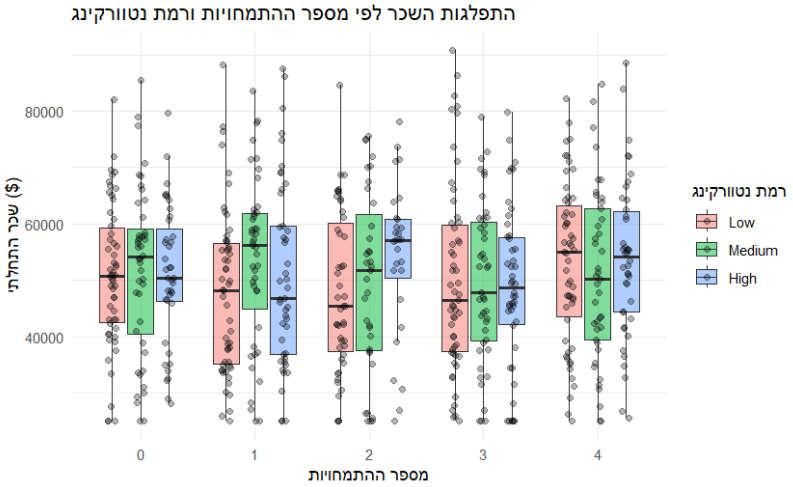
מדד התאמת המודל: (AIC) = 933.1

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

המשתנים אינם מנבאים מובהקים של הסיכוי להרוויח מעל חציון השכר. השפעתם קטנה מאוד ואינה מובהקת סטטיסטית. ($p > 0.05$)

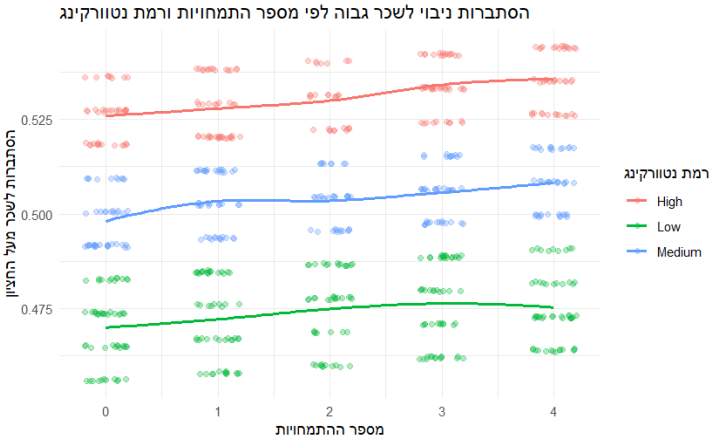
$Exp(\beta)$ קרוב ל-1, כלומר אין השפעה משמעותית של ההתמחויות או הנטוורקינג על רמות השכר.

גרף לתיאור כל אחד מהאפקטים שהתקבלו



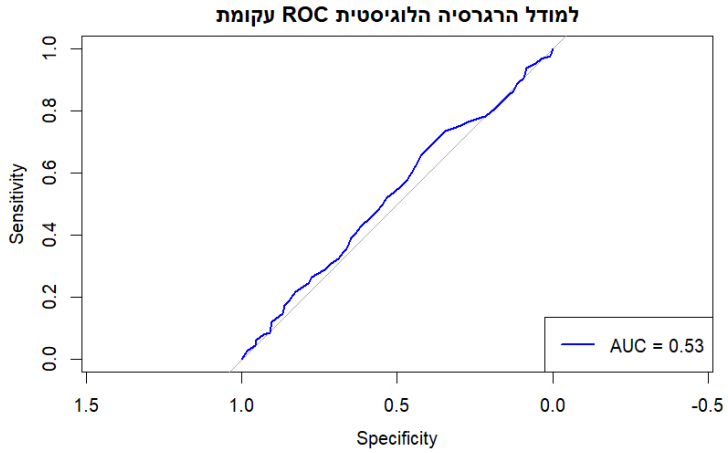
אפקט מספר ההתמחויות וכישורי הנטוורקינג על השכר
(עבור הרגרסיה הלינארית)

ניתוח הרגרסיה הלינארית הראה כי מספר ההתמחויות וכישורי הנטוורקינג אינם מנבאים משמעותית את השכר ההתחלתי. ($p > 0.05$)
הגרף תומך בממצא זה בכך שהוא מציג חפיפה משמעותית בין הקבוצות, חוסר דפוס ברור ושונות גבוהה בתוך כל קבוצה.
מסקנה: אין עדות להשפעה מובהקת של מספר ההתמחויות או רמת הנטוורקינג על השכר.



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

הגרף מציג את ההסתברות החזויה לשכר גבוה (מעל החציון) בהתאם למספר ההתמחויות (0–4) ולרמת הנטוורקינג (Low, Medium, High). ניתן לראות כי ההסתברות נעה בטווח שבין 0.47 ל-0.53, ללא מגמות חדות או שינויים משמעותיים. רמות הנטוורקינג אינן מראות הבדל ניכר בהשפעה על ההסתברות לשכר גבוה, ומספר ההתמחויות משפיע באופן מינימלי. ממצא זה תואם את תוצאות הרגרסיה הלוגיסטית, שבה שני המשתנים נמצאו כבלתי מובהקים סטטיסטית, ($p > 0.1$) מה שמעיד על תרומה נמוכה לניבוי השכר ההתחלתי.



גרף ROC עבור ניתוח הרגרסיה לוגיסטית

גרף ה ROC-מציג את יכולת המודל הלוגיסטי להבחין בין שכר גבוה מהחציון לשכר נמוך ממנו.

ערך AUC = 0.53, קרוב מאוד ל-0.5, מה שמעיד על ניבוי כמעט אקראי –כלומר, המודל לא מצליח להפריד בצורה טובה בין הקבוצות.
הקו הכחול בגרף כמעט חופף לקו האלכסוני (שמייצג ניחוש אקראי), דבר שמחזק את המסקנה כי מספר ההתמחויות וכישורי הנטוורקינג אינם מנבאים מהותיים לשכר גבוה.

מסקנה: המודל הלוגיסטי אינו מספק ניבוי משמעותי לשכר גבוה על בסיס משתנים אלו.