הערות על תרגיל טיטאניק

* המטרה המרכזית כאן היא שהנחפף יתחיל לעבוד עם גרפים, יבין איך אפשר להפוך שאלת מחקר על דאטאסט לגרף ברור ומובן, ויבין את ההבדל בין קורלציה לקוזאליות.
* אפשר לבדוק כאן הרבה קשרים בין כמעט כל אחד מהפרמטרים לסיכוי לשרוד, אבל החכמה היא להבין האם הקשר שאנחנו רואים הוא בגלל הפרמטר הנבדק, או האם יש פרמטר נסתר שמשפיע (לדוג' אם כבר הסקנו שנשים שרדו יותר מגברים, ועכשיו אנחנו רואים שרציף העלייה משפיע על הסיכוי לשרוד, חובה לבדוק את הסיכויים גם בהפרדה למין).
  + רעיונות לבדיקות – מין קורלטיבי (נשים שרדו יותר מגברים גם אחרי הפרדה לגילאים), ילדים שרדו יותר, יש קשר כלשהו גם למחלקה.
* צריך שיהיה הגיון מאחורי המסקנות (לדוג' – למה שרציף העלייה ישפיע? לא לזרוק פרמטרים סתם).
* לכן, בסוף הנחפף כנראה יגיע למצב בו הוא מפצל בין כמה פרמטרים ולכל חיתוך שלהם בודק את הסיכוי לשרוד. החכמה כאן היא **להציג את הנתונים בצורה קריאה**, לדוג' באמצעות scatter plot שבו אפשר לייצג כל אחד מהפרמטרים באמצעות ציר X, צבע, צורה, ועוד.
* דגשים קטנים:
  + גרפים מספיק גדולים עם כותרות לצירים.
  + לכל גרף חייבת להיות משמעות – "למה אתה מציג לי את הגרף הזה? מה אתה לומד ממנו?"
  + בהינתן הסיבה שלשמה צויר הגרף, הגרף צריך להיות ברור – התשובה לשאלת המחקר צריכה להיות יחסית ברורה בהינתן הגרף. אם צריך להשוות בין גרפים שונים בשביל להסיק מסקנה, או בין ערכים שונים שלא נוח להשוות ביניהם זה לא טוב.
  + סוג הגרף חשוב גם כן – צריך לראות שהנחפף לא עושה את אותו הגרף לאורך כל התרגיל, המטרה כאן היא גם לראות שהוא נחשף לגרפים מסוגים שונים ומתנסה בהם. מומלץ לומר לנחפף להסתכל על violin-plot, box-plot, histplot וכו'.
  + נרמול – בהרבה גרפים יש אפשרות לנרמל או לא לנרמל את ההתפלגות של הדברים המוצגים. לשים לב שהנחפף מנרמל כשצריך (לדוג' כשהוא משווה בין התפלגויות, וספציפית בהשוואות שהוא גם שם את ההתפלגויות על אותו הגרף).
  + כשמציגים אחוזים של סיכוי לשרוד מעניינת בחלק מהמקרים גם כמות האנשים עליהם האחוז חושב (שלא יצא מצב שהקבוצה קטנה ממש). אפשר להוסיף את זה לגרף (לדוג' ליד כל ערך בbarplot או scatterplot לכתוב את הכמות בנקודה).
  + המחברת צריכה להכיל רצף לוגי שהנחפף מציג, עם שאלות שהוא שואל כשבסוף (של המחברת או של הקטעים השונים) יש מסקנות ברורות.