****

תרגיל מסכם – **שיעור הכנת דשבורד**

שלום תלמידים,

תרגיל זה הוא **תרגיל מסכם** לכלל החומר שלמדתם בפרקי האקסל והסטטיסטיקה. על כן, לצורך פתרון תרגיל זה, אתם מוזמנים להשתמש בכל כלי שלמדתם עד כה, ושיכול לסייע לכם להגיע ליעד. אם אתם מכירים כלי אקסל וסטטיסטיקה נוספים שלא נלמדו בשיעורים, תוכלו להשתמש גם בהם אם הם מסייעים לכם.

**לצורך התרגיל,** עליכם לבחור באחד מבין מסדי הנתונים להלן**, אשר אותו אתם מעוניינים לחקור:**

[**סנאים בסנטרל פארק**](https://docs.google.com/spreadsheets/d/1XZvXPY62iDJSL8oWebJ86CMIn4XJ79Gd/edit#gid=496535465)

[**אייר בי.אנ.בי בוסטון**](https://docs.google.com/spreadsheets/d/1LcL2OB74KlKgWjVSGJpZeIZytTCVgN6T/edit#gid=1419677073)

[**כותרים בנטפליקס**](https://docs.google.com/spreadsheets/d/1d3cFWNBh0x6Dlnsos2s8rJ5cLI4geZJR/edit#gid=282871565)

[**גיבורי על**](https://docs.google.com/spreadsheets/d/1llIAXfmcmxBScm5-x5FSk_pOnIFxDZCO/edit#gid=879260660)

[**מדד האושר הבינלאומי**](https://docs.google.com/spreadsheets/d/1vZcEyceAlkftXM5gMNAnRRcPp4_hv_vM/edit#gid=1897893305)

**חלק 1:**

**הביטו בטבלה שבחרתם, ו**נסחו 3-5 שאלות חקר **עליהן תוכלו לענות באמצעות הנתונים אשר בידיכם. שימו לב שהשאלות מנוסחות כשאלות חקר, וכוללות הצגת משתנים תלויים ובלתי תלויים. השתדלו שהשאלות יעסקו בנושא כללי אחד בלבד.**

**חלק 2:**

**השתמשו בנתונים שבידיכם על מנת לתת מענה לשאלות החקר שניסחתם בסעיף הקודם.**

**נתחו את הנתונים, צרו טבלאות ציר ותרשימים מתאימים, ובסיום התהליך הפיקו דשבורד מלא. עליכם להפיק דשבורד אשר מעביר מסר. המסר צריך להיות כזה המבוסס על נתוני עבר, שלפיהם ניתן להפיק גם תובנות לעתיד.**

**את כל התהליך עליכם לתעד במסמך זה. מטה תוכלו לראות שאלות הכוונה לכל חלק בתהליך. תוכלו להשתמש בהן על מנת לתעד את תהליך ניתוח הנתונים, ותוכלו להוסיף גם הערות משלכם.**

**שאלות הכוונה לתיעוד תהליך ניתוח נתונים:**

1. **מדוע בחרתם במסד הנתונים הספציפי הזה?**

**בחרתי במסד הנתונים של נתוני הנוסעים על ספינת ה"טיטאניק". מעניין אותי להבין מה היה הסיכוי של כל נוסע לשרוד את הטביעה בהתחשב בנתונים שלו, ולראות מה עמד מאחורי קבלת ההחלטות של עובדי הספינה ברגעי החירום בהם הבינה כי הספינה עומדת לשקוע לקרקעית הים.**

1. **מהן השאלות עליהן תרצו לקבל מענה באמצעות הנתונים? (שאלות החקר)**

**מהו הסיכוי של נוסע על ספינת ה"טיטאניק" לשרוד את הטביעה בהתחשב בנתונים שלו?**

1. **אילו נתונים תצטרכו על מנת לענות על שאלות החקר?**

**נרצה להשתמש בנתונים האישיים של כל נוסע – מגדר, גיל (או קבוצת גיל), מעמד סוציו-אקונומי, מקום העלייה לספינה, וכמובן – נזדקק לנתוני השרידות על מנת להשוות ביניהם לבין שאר הנתונים. הייתי רוצה להשתמש גם בנתונים על עיסוק הנוסע (מקצוע), ובעיקר לדעת מי מהנוסעים היה עובד על ספינת הטיטאניק, אך נתונים אלה אינם קיימים במסד הנתונים הנחקר.**

1. **בדקו את הנתונים שבידיכם. האם ישנם נתונים חסרים/חריגים? אילו פעולות ניקוי וטיוב נתונים עליכם לעשות על מנת להתחיל לעבד את הנתונים?**

**הסרתי את שורות 17 ו-116 שהיו ריקות.**

**הסרתי עמודות שאינן רלוונטיות למחקר – עמודת Name, עמודת Ticket, עמודת Cabin, עמודת Boat, עמודת Body, עמודת Home.dest.**

**הפכתי את טווח הנתונים לטבלה מעוצבת לשם נוחות העבודה באקסל. באמצעות כפתור הסינון, בדקתי אם קיימים ערכים חריגים/חסרים בכל עמודה.**

**ראיתי שבעמודת Age חסרים ערכים רבים (263 ערכים מתוך 1309, שהם כ-20%). במקרה זה יש שתי אפשרויות – נוכל לבחור להסיר את העמודה ולא להשתמש בה כלל עקב החוסרים הרבים (החוסרים עלולים להוות גורם מטעה בבדיקת מתאם המשתנים), או שנוכל להמשיך ולבדוק מאפיין זה תוך התעלמות מהחוסרים. בחרתי להמשיך בעבודה עם מאפיין Age תוך התעלמות מהחוסרים.**

**ישנם שני ערכים חסרים בעמודה Embarked (שמשמעותה – מקום העלייה לספינה). בחרתי להציב בהם את הנתון השכיח ביותר בעמודה, והוא האות S (משמעותה שמקום העלייה הוא העיר סאות'המפטון, אנגליה).**

1. **אילו פעולות ביצעתם במהלך ניתוח הנתונים? באילו כלי אקסל וסטטיסטיקה השתמשתם?**

**לבדיקות הנתונים השתמשתי** בחישובי שכיחות **וב**טבלת שכיחויות **על כל העמודות, על מנת לזהות ערכים חריגים/חסרים.**

**ערכתי** טרנספורמציה **למספר עמודות:**

1. **את הערכים בעמודה Sex בחרתי להחליף ל-0/1 (0=גבר, 1=אישה), לשם נוחות העבודה באקסל. עשיתי זאת באמצעות שימוש בפונקציית IF – אם הערך הוא male, נחליף אותו ב-0, אחרת נחליף אותו ב-1.**
2. **עקב השונות הרבה של הערכים בעמודה Age, בחרתי להחליף אותם בערכים מקובצים ל-5 קבוצות גיל:**
3. **0-15**
4. **16-28**
5. **29-40**
6. **45-65**
7. **65+**

**גם טרנפורמציה זו נעשתה באמצעות פונקציית If, תוך שימוש** בפונקציות מקוננות**. במקום הערכים הריקים הוכנס הסימן "-".**

**השתמשתי ב**היסטוגרמה **כדי לזהות התפלגות ערכים של מספר עמודות (Age (לפני ואחרי הטרנספורמציה), Pclass, Sex).**

**השתמשתי בכלי** הטבלה המעוצבת **באקסל לשם נוחות העבודה.**

**לאחר מכן יצרתי מספר** טבלאות ציר **לסכימת הנתונים ולצורך יצירת גרפים.**

1. **מהם תוצרי ניתוח הנתונים? (נתוני סיכום, טבלאות ציר וכו')**

**הנתונים סוכמו במספר טבלאות ציר:**

1. **טבלה 1 – ספירת הנוסעים לפי נתוני שרידות (שרד / לא שרד).**
2. **טבלה 2 – ספירת הנוסעים לפי מגדר.**
3. **טבלה 3 – ספירת הנוסעים לפי מגדר + שרידות.**
4. **טבלה 4 – ספירת הנוסעים לפי מעמד סוציו-אקונומי + שרידות.**
5. **טבלה 5 – ספירת הנוסעים לפי גיל + שרידות. הטבלה מציגה את הנתונים באחוזים מתוך כל קבוצת גיל.**
6. **טבלה 6 – ספירת הנוסעים לפי מקום העלייה לספינה + שרידות. הטבלה מציגה את הנתונים באחוזים מתוך כל מקום עלייה.**
7. **באילו גרפים/תרשימים בחרתם להשתמש על מנת להציג את התוצרים? מדוע?**

**עבור כל הגרפים נבחר עיצוב תלת-מימדי, אשר מעניק נפח ו"חיים" לדשבורד. בדשבורד הזה יש להתבונן בנתונים באופן השוואתי (ולא באופן נומינלי, כלומר – המספרים לבדם לא בעלי חשיבות גבוהה), על כן לא קיימת בעיה בשימוש בגרפים תלת-מימדיים.**

1. **טבלה 1 – בחרתי להשתמש בגרף עוגה תלת-מימדי. בחרתי בגרף זה על אף הבעייתיות בתפיסה התלת-מימדית, כיוון שהוא מציג שני ערכים בלבד (שרד / לא שרד), ובמצב של שני ערכים אין בעיה בתפיסה התלת-מימדית של הגרף.**
2. **טבלה 2 – מאפיין המגדר בפני עצמו (ללא נתוני השרידות) אינו מעניק לנו זווית התבוננות נוספת על הנתונים, לכן לא הוגדר עבורה גרף. טבלה זו תשמש עבור נתונים סטטיים בדשבורד בלבד.**
3. **טבלה 3 – בחרתי להציג את נתוני הטבלה בגרף עמודות תלת-מימדי. הגרף נחוץ לצורך השוואתי (גברים מול נשים ונתוני שרידות), לכן לא קיימת בעיה בשימוש בגרף התלת-מימדי, כאשר רואים את כל הנתונים זה לצד זה וקל להשוות ביניהם.**
4. **טבלה 4 – גם כאן בחרתי להשתמש בגרף עמודות תלת-מימדי, בדומה לגרף מטבלה 3.**
5. **טבלה 5 – גם כאן בחרתי להשתמש בגרף עמודות תלת-מימדי, בדומה לגרף מטבלה 3.**
6. **טבלה 6 – בחרתי להשתמש בגרף ברים (אופקי) תלת-מימדי מוערם. בשימוש בגרף מוערם ניתן לראות גם את ההתפלגות הכללית של קבוצות הגיל בין נוסעי הספינה. הבחירה בגרף ברים היא כיוון שכך אוכל לסדר את הדשבורד באופן שנראה טוב לעין.**
7. **פרטו על בחירותיכם בתהליך הכנת הדשבורד – מה עומד מאחורי בחירת סידור האובייקטים? באילו צבעים בחרתם ומדוע? אם בחרתם להשתמש באלמנטים ייחודיים – הסבירו מדוע בחרתם בשימוש בהם, ומה הם מוסיפים לדשבורד.**

**בחרתי להשתמש** בכל הגרפים **בשני צבעים – צבע כחול עבור נוסעים ששרדו, וצבע אפור עבור נוסעים שלא שרדו. כחול הוא צבע רגוע שמשדר "חיות", לכן קל לקשר אותו עם שרידות. אפור הוא צבע "חסר חיים", לכן קל לקשר אותו עם אי-שרידות, וכך הצופה בדשבורד לא צריך להתאמץ כדי להבין את הנתונים בהם הוא מתבונן.**

**ניתן היה גם להשתמש בצבעי אדום וירוק (ירוק=שרד, אדום=לא שרד), אך בחרתי שלא לעשות זאת, כיוון שאדום הוא צבע מאוד בולט לעין, וגם מקושר במוח לאזהרה, כך שאילו הייתי בוחרת להשתמש בצבע אדום – היה קשה יותר להתבונן בגרפים שבדשבורד.**

**הוספתי נתונים סטטיים על מספר הנוסעים הכללי על הספינה, מספר הגברים ומספר הנשים. הוספתי גם אייקונים לצד נתונים אלה, על מנת שיהיה קל יותר לקרוא אותם, ובכדי להוסיף קצת "צבע" לדשבורד.**

**סידרתי את הדשבורד באופן שיהיה קל ונוח להבנה, תוך מחשבה על "זרימת" הנתונים:**

**בצד שמאל מיקמתי את כלי הפריסה, המקושרים כולם לכל הגרפים.**

**הגרף הראשון הוא גרף העוגה, המציג נתוני שרידות בלבד. זהו הגרף הפשוט ביותר, וגם החשוב ביותר, שכן כאשר פורסים את הנתונים – ניתן להסתכל עליו ישירות ולהבין בקלות כמה שרדו מתוך קבוצת הנוסעים שנבחרה בפריסה.**

**אחריו מגיעים גרפים המציגים את השרידות לפי מאפיינים נוספים. הסדר נקבע לפי מספר הערכים של כל מאפיין, מהקטן לגדול:**

1. **גרף שרידות לפי מגדר (2 ערכים – גבר/אישה)**
2. **גרף שרידות לפי מעמד (3 ערכים – 1/2/3)**
3. **גרף שרידות לפי מקום העלייה לספינה (3 ערכים – S/Q/C)**
4. **גרף שרידות לפי גיל (6 ערכים – 5 קבוצות גיל + שדות ריקים).**
5. **פרטו את מסקנותיכם מתהליך ניתוח הנתונים (מסקנות = דברים שניתן להסיק מנתוני העבר).**

**המסקנות שלי מתהליך ניתוח הנתונים:**

1. **ניתן לראות בדשבורד את קבלת ההחלטות בפועל – ראשונים לרדת בסירות ההצלה היו נשים, ילדים, ואנשי המעמד הגבוה (P\_Class 1). רואים זאת כאשר מסתכלים על סיכויי ההישרדות לפי מגדר, גיל, ומעמד – אלו הגורמים המשפיעים ביותר על סיכויי ההישרדות של הנוסעים.**
2. **אין משמעות ממשית לחקירת סיכויי השרידות לפי מקום העלייה לספינה. ניתן לראות כי סיכויי ההישרדות גבוהים יחסית עבור נוסעים שעלו בתחנת Cherbourg, אך בהתבוננות מעמיקה יותר מגלים כי הסיבה לכך היא שמרבית הנוסעים מתחנה זו הם מ-P\_Class 1, וזו הסיבה האמיתית לשרידותם.**
3. **סיכויי ההישרדות הגבוהים ביותר היו עבור ילדות ונשים מבוגרות (65+) בנות המעמד הגבוה – מתוך 15 נוסעים בקטגוריה זו, רק אחת לא שרדה.**
4. **סיכויי ההישרדות הנמוכים ביותר היו עבור גברים מבוגרים (46+) בני המעמד הנמוך – מתוך 16 נוסעים בקטגוריה זו, אף אחד לא שרד.**
5. **פרטו את תובנותיכם והמלצותיכם מתהליך ניתוח הנתונים (תובנות = השלכת המסקנות על העתיד / כיצד ניתן לשפר את ההתנהגות וקבלת ההחלטות בעתיד).**

**סיכויי השרידות הכלליים על אוניית הטיטאניק היו 38%. עבורי – נערה בכיתה י"א ממעמד סוציו-אקונומי בינוני – הסיכויים לשרוד הם 86%. אמנם לא מדובר בסיכוי נמוך במיוחד, אך יש תמיד לשאוף ל-100% (בהנחה שהדבר אפשרי כמובן).**

**ע"פ ניתוח הנתונים, על מנת שלא יתרחש שוב אסון בסדר גודל שכזה שבו סיכויי ההישרדות של נוסע הם 38%, מומלץ לוודא שקיים מספר סביר של סירות הצלה, ולא להסתמך על בניית ספינה "שאינה ניתנת לטביעה".**