**תרגיל מעשי:**

כחלק מהמאמץ הלאומי במלחמה בקורונה משרד הבריאות והמל"ל פנו לתוכנית פסגות בבקשה שיעזרו להם להעריך מה הגורמים לחולי קשה בעקבות מחלת הקורונה.

משרד הבריאות סיפק לכם מידע על 41580 חולי קורונה מאומתים.

להלן פירוט המידע המסופק עבור כל חולה קורונה מאומת:

1. age – גילו של החולה.
2. background\_diseases: **כל** מחלות הרקע המאובחנות של החולה.
3. background\_diseases\_binary: האם קיימות מחלות רקע (משתנה הנגזר מהמשתנה הקודם).
4. confirmed\_date: תאריך אבחון של מחלת הקורונה.
5. date\_onset\_symptoms: תאריך הופעת סימפטומים של מחלת הקורונה.
6. deceased\_date: תאריך פטירה (אם קיים).
7. severity\_illness: חומרת המחלה.
8. sex: מין החולה.
9. smoking: האם החולה מעשן.
10. treatment: הטיפול אותו מקבל החולה.

המשימה:

הגדרת הבעיה המדוייקת קשה למשרד הבריאות ולכן נחלק אותה למספר שאלות:

1. איזה סוג אנשים צפויים למות?
2. איזה סוג אנשים צפויים לחלות בצורה קשה?
3. איזה סוג אנשים צפויים להצטרך לאישפוז?

יש לענות על שאלות אלו באמצעות נתונים מספריים וויזואליזציות ברורים ומוסברים היטב.

דגשים והערות:

1. כמו בכל פרויקט, המידע הגולמי שמתקבל מהלקוח עשוי להיות מלוכלך ולהכיל הרבה בעיות. לכן עליכם להיות חדים, לחוש את המידע ולחפש באופן שיטתי אחר בעיות אפשריות במידע.
2. תעד את כל הבעיות במידע בהן נתקלת במסמך וורד נפרד. הרבה פעמים, ריכוז הבעיות במידע עבור הלקוח הן תוצר מעניין בפני עצמו של הפרוייקט. כאשר אתה מתעד בעיות עשה זאת באופן כמה שיותר מסודר וברור, כך שהלקוח יוכל לצפות בהן בעצמו בקלות.
3. אל תהססו לבקש מהלקוח (קרי החונך) הסברים למידע שאינכם מבינים, או לבעיות במידע שמצריכות פירוש/הבהרה.

**מדריך לחונך:**

הישגים נדרשים:

ראשית, במעבר על התרגיל יש להסביר שהבעיה לא מוגדרת היטב, וזה לא מקרה נדיר.

1. ניקוי המידע, בדיקת הטיות וערכים חסרים כפי שיפרוט בהמשך.
2. ויזואליזציות, קורלציות, בדיקת השערות ונתונים אינפורמטיביים נוספים עבור החלוקות הבאות, בנפרד וביחד:
3. לפי מחלות רקע:

ראשית לפי האם קיימת מחלת רקע כלשהי.  
פר מחלת רקע (יש בסך הכל 8 מחלות רקע שונות. שדה זה הוא מחרוזת מופרדת בפסיקים ולכן מומלץ להפריד ל8 שדות בוליאניים.).

1. לפי גיל:

רצוי להסתכל גם על קבוצות גיל.

1. לפי מין.
2. לפי עישון.
3. ימים מרגע הופעת הסימפטומים ועד אבחון.

כחלק מהאקספלורציה נדרש להבחין בבעיות ובחוסרים הבאים העולים מן המידע:

1. חניך: המשתנה background\_diseases\_binary לא באמת "נגזר" באופן ישיר מהמשתנה background\_diseases.

דהיינו, יתכנו חולי קורונה אשר יש להם מחלות רקע והמשתנה הבינארי שלהם הוא 0 ולהיפך.  
חונך: המשתנה האמין ביותר הוא background\_diseases ולכן יש לפעול לפיו.

1. חניך: במשתנה sex יש יותר משני ערכים אפשריים:

female, male, man, woman, boy, girl, unknown

חונך: צריך להביא את זה לצורה מוסכמת של שניים/ שלושה ערכים.

1. חניך: ישנם חולים שקיים עבורם deceased\_date אולם ה- severity\_illness ריק.

חונך: אם קיים deceased\_date החולה אכן מת.

1. חניך: ישנם חולים שהערך smoking חסר אצלם.

חונך: להפנות את תשומת הלב להתפלגות המעשנים באוכלוסיית החולים.   
ההתפלגות מראה כי יש פחות מ10% מעשנים ולכן כדאי להשלים ערך זה באמצעות קביעה כי כל מי שערך זה לא נתון לגביו לא מעשן.  
דרכים נוספות: קיימים דפוסי עישון שונים בין גברים ונשים לכן לדגום לפי מין.  
קיימים דפוסי עישון שונים לפי קבוצות גיל ולכן לדגום לפי קבוצת גיל.

1. חניך: ישנם חולים שה- severity\_illness הוא קשה או מת אך ה- treatment שלו הוא מרפאה קהילתית.

חונך: כל מי שחולה קשה או מת אושפז.

1. חניך: ישנם חולים עם גיל 119.

חונך: אלו אנשים מאוד מבוגרים שגילם המדוייק לא ידוע. לכן (אם עדיין לא הגיעו לכך אז להפנות את תשומת ליבם) בחלוקה של הגיל לקבוצות יש לשייכם לקבוצה המבוגרת ביותר.

1. חניך: ישנם חולים ללא גיל.

חונך: יש לשערך את הגיל.

דרכים לפתרון: ממוצע/חציון.

מחלות רקע בד"כ באות עם הגיל ולכן לדגום את הגיל לפי קבוצת בעלי המחלות רקע הזהות.

1. חניך: ישנם חולים שנדבקו לפני נובמבר 2019.

חונך: הרישום הוא ידני וזו טעות ברישום.  
מדובר באותו יום וחודש אבל בשנת 2020.

1. חניך: ישנם חולים שתאריך הופעת הסימפטומים הוא לאחר תאריך האימות כחולים.

חונך: הגיוני, אנשים שנבדקו והיו א-סימפטומטיים והתגלו כחיוביים והופיעו הסימפטומים לאחר מכן.

1. חניך: ישנם חולים ללא תאריך הופעת סימפטומים.

חונך: חולים א-סימפטומטיים.