תרגיל בית 2 - **Data science**

מועד הגשה: 5.5 בשעה 23:55

הנחיות כלליות להגשת תרגילי הבית:

- הגשת התרגילים תתבצע באמצעות אתר הקורס בלבד.
- יש להגיש את התרגיל לתיבת ההגשה המתאימה, בהתאם לפורמט הבא: HW#_ID1_ID2 , כאשר התו # מציין את מספר התרגיל,ID1 מציין את תעודת הזהות של הסטודנט הראשון, ID2 מציין את תעודת הזהות של הסטודנט השני.

לדוגמה: HW1_000000000_111111111.py זהו קובץ בודד שפותר את תרגיל בית 1.

- בנוסף, יש לכתוב כהערה (בתוך הקוד) בראש כל קובץ את השמות המלאים (שם פרטי ושם משפחה) בלועזית ואת תעודת הזהות של כל אחד מהשותפים, כאשר הפרטים של כל אחד רשומים בשורה נפרדת.
- ההגשה מתבצעת באמצעות קבוצות ההגשה בלבד.
- תרגיל אשר יוגש באיחור לא ייבדק, למעט אישורים מיוחדים כגון: מילואים, אשפוז וכו' המתקבלים לפני מועד ההגשה הרשמי.
- יש לכתוב הסברים ומענה מילולי על השאלות בגוף הקובץ בעזרת הערות ובאנגלית בלבד.
- יש להפריד בין פתרונות של כל שאלה בצורה הבאה. לפני המענה לשאלה מספר 1 כתבו בהערה Q1 וכן הלאה. במידה והשאלה מכילה מספר סעיפים (נניח 2 סעיפים) יש לכתוב בהערה Q1a לפני פתרון סעיף a בשאלה 1, ו- Q1b לפני פתרון סעיף b בשאלה 1 וכן הלאה.

בהצלחה!

מצורף קובץ csv **המכיל נתונים על 198 אורחי קלאב מד, ששהו במלון בשנת 2016.**

שימו לב! בשאלות הבאות, כאשר אתם מבצעים שינוי בדאטה יש לשמור את הערכים לאחר השינוי במשתנה חדש (עמודה חדשה) ולא לדרוס את הנתונים המקוריים.

1) א) הציגו את גילאי הלקוחות באמצעות היסטוגרמה ותנו שמות לצירים וכותרת.
ב) הדפיסו עבור חלוקות שונות (גדולות וקטנות יותר) של bins , מה ניתן ללמוד מכל אחת מהן?

2) הציגו את נתוני הפרמטר club_member בגרף מתאים לדעתכם. הוסיפו כותרת לגרף, שמות לצירים וצבע.

3) מצאו עמודה נומרית בעלת התפלגות מוטה (ימינה/שמאלה) והפעילו עליה טרנספורמצית log. הסבירו איך הטרנספורמציה השפיעה על ההתפלגות, האם עזרה?

4) צרו טבלאות פרופורציה בשני אופנים עבור המשתנים הקטגורים : סטטוס משפחתי ומין.
א) טבלה עבור התפלגות הסטטוס המשפחתי לפי מין.
ב) טבלה עבור התפלגות מין לפי סטטוס משפחתי.
ג) יש לצייר עבור כל טבלה גרף עם צבעים ומקרא .

היעזרו בסעיפים א' ו-ב' כדי לענות על השאלות הבאות:

- באיזה סטטוס משפחתי אחוז הגברים הינו הגדול ביותר? האם בהכרח כמות הגברים באותו הסטטוס הינה הגדולה ביותר מבין כל הסטטוסים המשפחתיים?
- מהו הסטטוס המשפחתי השכיח ביותר בקרב הנשים?
- מה אחוז הנשואות מכלל הנשים ?
- מהו אחוז הגברים מכלל הרווקים/ות?

ד) צרו טבלת פרופורציות כרצונכם מ 2 משתנים מתאימים והציגו אותה. תנו לה כותרת המתארת את הצגת הנתונים והוסיפו צבעים ומקרא. הסבירו מה ניתן ללמוד ממנה.
ה) איזה מבין המשתנים הבאים מתואם יותר (כלומר: יש לו קשר מגמתי) למשתנה מין:
club_member או status ?

5) שרטטו גרף מתאים של הוצאת המיניבר כנגד גיל האורחים. הוסיפו כותרת לגרף ושמות לצירים.

6) א. חשבו את הרבעונים (quartiles) והדפיסו את הטווח הבין-רבעוני (interquartile range) ואת סטיית התקן של המשתנה room_price (שימו לב לא לכלול ערכים חסרים).
ב. בדקו כמה ערכים קיימים בדאטה שקטנים או שווים לחציון של המשתנה room_price. האם זה תואם את ההגדרה של החציון? אם לא אז הסבירו ממה נובע ההבדל.
ג. שרטטו על היסטוגרמה את הנתונים והוסיפו 3 קווים אנכיים ב 2 צבעים שונים- אחד עבור הממוצע ושניים עבור טווח סטיית תקן אחת.
ד. עבור המשתנה room_price האם ההתפלגות היא בקירוב נורמלית? האם היא רחבה /צרה / זהה להתפלגות הנורמלית, מוטת ימינה/סימטרית/מוטת שמאלה?

ה. שרטטו multivariate boxplot של התפלגות הגילאים - לפי דירוג האורחים (המשתנה ranking).
עפ״י הנתונים בתרשים:

א. עבור איזה מהדרגים (ranking) הטווח הבין-רבעוני מתפרש על פני טווח הגילאים הרחב ביותר?

ב. השתמשו בפונקציית השרטוט axhline והוסיפו לשרטוט של א' קו אופקי שמפריד את גילאי השוליים (outliers).

ו. שרטטו multivariate boxplot של התפלגות הגילאים לפי תדירות הביקורים ב 5 השנים האחרונות (המשתנה visits5years). מהו מספר הביקורים עבורו הטווח הבין-רבעוני מתאים לאוכלוסיה המבוגרת ביותר? איזה גילאים נכללים בקבוצה זו ?

ז. מה תוכלו לומר על קטגוריית מספר הביקורים שמצאתם בסעיף ו' בהשוואה לקטגוריות הביקורים האחרות, ביחס להוצאת מחיר החדר (המשתנה room price) ?

ח. האם ניתן לראות קשר מגמתי מובהק בין הדירוג של אורחי המלון לסך כל ההוצאות שלהם במלון ? יש לנמק בעזרת תרשים מתאים ולהסביר בקצרה.

7. סדרו מחדש את המשתנה 2016visits (מס הביקורים הכולל בשנת 2016) :
הפכו את המשתנה לקטגורי בעל 2 קטגוריות בלבד ותנו שמות אינפורמטיבים (בעלי משמעות) לקטגוריות החדשות. עבור חלק מהלקוחות חסרים נתונים כיוון שהם הצטרפו רק לאחר 2016. יש לתת קטגוריה גם עבורם.
הדפיסו תיאור עמודה בסיום.

8. א. המשתנה total_expenditure מסמל את <u>סה״כ ההוצאה הכוללת של אורחי קלאב מד בסוף שהייתם במלון</u>.
עזרו להנהלת המלון לדרג את לקוחותיה לפי סה״כ ההוצאות שלהם באופן הבא:

- תחילה, יש לטפל בנתונים שגויים (לא הגיוניים) כך שהממוצע של הערכים התקינים לא ישתנה.
- לאחר מכן, יש לבצע חלוקה בדידה (discretization) של ההוצאות לפי רבעונים (quartiles) ולשמור זאת בעמודה חדשה בשם "total_expenditure_new".
  שימו לב! לטווחים בחלוקה לרבעונים (כדאי להיעזר בפונקציה unique).
- לבסוף, הגדירו שמות לקטגוריות כך שיתארו את החלוקה.

ב. כיצד לדעתכם כדאי לטפל בערכים שגויים ו/או חסרים בעמודת room_price? האם באמצעות החלפה בחציון או בממוצע? מדוע?

ג. חזרו על סעיף א' רק הפעם עבור חלוקה לפי 3 סטיות תקן סביב הממוצע (ס״הכ 6 קטגוריות). כמה אורחים לא יכללו בחלוקה זו?
לדוגמה: עבור סטיית תקן 1 וממוצע 0 נקבל את הקטגוריות הבאות.
קטגוריה 1 - כל המספרים בין 0 ל-1.
קטגוריה 2 - כל המספרים בין 0 ל (-1).
קטגוריה 3 - כל המספרים בין 1 ל-2.
קטגוריה 4 - כל המספרים בין (-1) ל (-2).
קטגוריה 5 - כל המספרים בין 2 ל-3.
קטגוריה 6 - כל המספרים בין (-2) ל (-3).

9. א. בצעו נרמול לפרמטר minibar.
ב. הדפיסו את סטיית התקן לפני נרמול ואחרי.
ג. הדפיסו גם את כמות הערכים הטיפוסיים.