עבודת סיכום מס' 4

נושאים:

1. *פרוטוקול שליפה*
2. *בדיקת הנתונים (EDA)*
3. *ניקוי הנתונים*

הערה: ניתן לעבוד ב-Python או ב-R. מומלץ לעבוד ב-jupyter notebook. מי שבוחר לעבוד ב-RStudio, בנוסף לקוד יצטרך לכתוב את התובנות ולהעתיק את הגרפים במסמך word.

חלק 1 - פרוטוקול שליפה

1. להוריד את פרוטוקול השליפה מהלינק ולהשלים את כל העמודות. קוד ה-SQL כדי לייצר את ה-VIEW עם הקובץ השטוח נמצא בלינק.

חלק 2 - בדיקת הנתונים - Exploratory data analysis

1. תתארו את הנתונים עם סטטיסטיקה תיאורית (תשתמשו במדדי מירכוז ופיזור)
2. תייצרו גרפים המתארים את ההתנהגות של כל משתנה.
3. תייצרו מטריצה קורלציות והציגו אותה בגרף
4. תתארו את משתנה המטרה (revenue) - איך הוא מתפלג? האם יש קטגוריות שמראים שוני גדול בהתפלגות של משתנה המטרה?
5. תייצרו גרפים שיכולים לעזור לכם לבדוק האם קיימים נתוני קיצון. תתארו אותם.
6. תתארו את הנתונים החסרים: אצל איזה משתנים יש נתונים חסרים? כמה?
7. תייצרו מטריצה של חסרים (תייצרו dataframe עם אותם מימדים מטבלת המקור ותאים שיש חסרים תשימו ערך של אחד ובאלה שיש נתונים ערך אפס). תציגו את המטריצה בגרף heatmap.

חלק 3 - ניקוי הנתונים

3.1 נתוני קיצון

1. במשתנים שבהם מוצאים ערכי קיצון תבדקו את ההתפלגות של המשתנה עם ובלי ערכי הקיצון. האם ההתפלגות משתנה ?
2. תייצרו גרף scatter עם המשתנים שבשאלה הקודמת ב-X ומשתנה המטרה ב-Y. האם ערכי קיצון במשתנה ה-X משפיעה על ההתנהגות של משתנה ה-Y? האם רואים שוני עם או בלי ערכי הקיצון?
3. באיזה משתנים הייתם מוחקים את ערכי הקיצון? איך הייתם מוחקים אותם? נמקו.
4. תפעלו על הנתונים לפי מה שהגדרתם בשאלה הקודמת.

3.2 נתונים חסרים

1. עבור כל משתנה עם נתונים חסרים, תראו את ההתפלגות **של משתנים אחרים** עם או בלי חסרים. השתמשו במטריצת החסרים שייצרתם בחלק 2, שאלה 7 עבור החיווי של יש/אין חסר. עבור ההתפלגות, תשתמשו בהיסטוגרמה או בגרף density עם קטגוריה/צבע לפי החיווי.
2. תייצרו טבלה של המשתנים שבהם יש חסרים ותתארו מהו מנגנון היצירה של החסרים (מבוסס על התוצאות של השאלה הקודמת).
3. איזה טכניקה imputation מתאימה לכל משתנה? תשתמשו בטכניקה הנבחרת כדי להחליף את החסרים.

חלק 4 - אחרי ניקוי הנתונים, תחזרו על החלק 2 במלואו.