**דרישות פרויקט גמר קורס –התמחות Data Science**

נתון קובץ (מצורף למטלה) עם נתוני מחירים של בתי מלון בניו יורק. המחירים לבתי המלון השונים התקבלו במספר דגימות, בכל דגימה התקבלו מחירים לבתי מלון בתאריכים שונים, עד חודש מתאריך הדגימה. שימו לב כי לא מובטח כי בכל דגימה יופיעו כל בתי המלון. הקובץ מכיל את השדות הבאים:

* 1. Snapshot – מזהה הדגימה אשר בה ביקשנו את המחירים (ניתן להתעלם מנתון זה)
  2. SnapShot Date - תאריך הדגימה
  3. Checkin Date – תאריך הכניסה למלון שמעוניינים בו
  4. Days - משך השהיה במלון בימים – שימו לב שקיבלתם רק נתונים לשהיה של 5 ימים.
  5. Original Price – מחיר החדר ללא הנחה מיוחדת - בדולרים
  6. Discount Price – מחיר החדר לאחר הנחה (ההנחה תלויה גם בסוג קוד ההנחה) - בדולרים
  7. Discount Code – קוד הנחה (מספר בין 1 ל 4) – המשפיע על גובה ההנחה
  8. Available Rooms – מספר החדרים הזמינים בתאריך זה (שימו לב כי -1 מציין כי לא ניתן היה להשיג מידע על מספר החדרים)
  9. Hotel Name – שם בית המלון
  10. Hotel Stars – מספר הכוכבים של בית המלון

משימות לביצוע:

1. **הוספת מאפיינים**

באמצעות קוד פייתון ושימוש ב Pandas, עליכם להוסיף לקובץ הנתונים את העמודות הבאות:

* 1. DayDiff - מספר הימים בין ה Snapshot Date לבין ה Checkin Date
  2. WeekDay - היום בשבוע (Sun/Mon/Tue/Wed/Thu/Fri/Sat) של ה Checkin Date
  3. DiscountDiff - גובה ההנחה בדולרים (ההפרש בין Original Price ל Discount Price)
  4. DiscountPerc - גובה ההנחה באחוזים

הקובץ לאחר השינויים ייקרא בשם : Hotels\_data\_Changed.csv. מעתה והלאה – כל החישובים יבוצעו על סמך הקובץ החדש שיצרתם

1. **למידת קוד ההנחה הזול ביותר באמצעות Classification**
   * 1. נרצה לבדוק אם ניתן לחזות את קוד ההנחה הטוב ביותר מבין ארבעת הקודים האפשריים לפי יתר הפרמטרים. על מנת לבדוק זאת, עליכם ליצור קובץ המאפשר עבור כל:

Snapshot Date+ Checkin Date + DayDiff + Hotel Name + WeekDay מהו קוד ההנחה המקנה את ההנחה המקסימלית (קוד ההנחה יהיה ה Class שלנו).

* + 1. לאחר יצירת הקובץ החדש, נרצה להריץ אלגוריתמי קלאסיפיקציה באמצעות Python אשר יחזה את קוד ההנחה המקסימאלית בהינתן הפרמטרים שהזנו. הריצו על 2 אלגוריתמי קלאסיפיקציה (אחד על בסיס עץ החלטה, והשני על בסיס Naïve Bayes)
    2. הציגו Screenshots של תוצאת הבדיקה, והסבירו את משמעות התוצאות שקיבלתם. השוו בין 2 האלגוריתמים ודונו על ההבדלים ביניהם.
    3. **הסבירו מהם הפרמטרים שהשפיעו הכי הרבה** על ההחלטה ונסו לכמת עד כמה כל פרמטר משפיע.
    4. הראו באמצעות **מדדים סטטיסטיים** את איכות התוצאות (ROC, FP, FN).

1. **למידת אשכולות של בתי מלון על פי מדיניות התמחור באמצעות ביצוע Clustering**

נרצה למצוא אשכולות של בתי מלון המכילים התנהגות תמחור דומה לאורך התקופה בה התקבלו הנתונים. על מנת לבצע זאת, נרצה לשנות את פורמט הנתונים כך שלכל בית מלון (לפי שם בית המלון) נבנה וקטור המתאר את המחיר המנורמל (לאותו תאריך), לאורך התקופה.התהליך אותו נבצע:

* + 1. מצא את 150 בתי המלון המכילים הכי הרבה רשומות בקובץ המקורי. ניקח רק נתונים של מלונות אלה
    2. מצא את 40 תאריכי ה Checkin המכילים הכי הרבה רשומות. ניקח רק את הנתונים של תאריכים אלה.
    3. עבור כל תאריך Checkin, ניקח את 4 המחירים עבור 4 קודי ההנחות. – סה"כ נקבל לכל בית מלון 160 מספרים (40 תאריכים \* 4 מחירים). שימו לב שייתכן כי בתאריכים מסוימים לא יימצאו נתונים עבור בית המלון הספציפי – במקרה זה נסמן מחיר של -1. שימו לב כי כל לתאריך Checkin יש יותר מ snapshot אחד, עבור כל תאריך Checkin לבית מלון מסוים ניקח את זה בעל המחיר (לאחר הנחה) המינימאלי מבין כל ה snapshot-ים של מלון זה לאותו תאריך Checkin
    4. את המחירים יש לנרמל בין 0 ל 100 – כך ש 0 זה המחיר הנמוך ביותר שמצאנו מתוך 160 המחירים (כמובן לא כולל -1), ו 100 הוא המחיר הגבוה ביותר שמצאנו עבור בית המלון הספציפי
    5. קיבלנו קובץ שבו כל שורה מייצגת בית מלון אחד, ויש בה 161 עמודות – אחת לשם המלון ו 160 עבור המחירים המנורמלים
    6. על הקובץ שיצרנו נריץ אלגוריתם Clustering (מסוג Hierarchical Clustering) באמצעות Python . יש להציג Dendrogram (עץ היררכיות) המתאר את ה Cluster-ים של בתי המלון בעלי מדיניות התמחור הדומה. הציגו Screenshot של תוצאת הבדיקה, והסבירו את משמעות התוצאות שקיבלתם.

1. **מימוש השלבים 2,3 באמצעות SPARK**

עליכם לממש באמצעות PySpark (שילוב Spark עם פייתון) וספריית אלגוריתמי הלמידה MLLib את השלבים 2,3 – קלאסיפיקציה (שאלה 2 במסמך זה) או קלאסטרינג (שאלה 3 במסמך זה) לבחירתכם.