הפקולטה למנהל עסקים - מסלול למערכות מידע

ניתוח נתונים באמצעות PYTHON – תרגיל הגשה 1

הוראות כלליות:

יש להגיש לאתר הקורס קובץ ipynb מלא על בסיס הקובץ Python_data_analysis_Ex1.ipynb בו מופיעות שאלות תרגיל זה.

שימו לב! לתרגיל שאלות מילוליות בנוסף למקטעי הקוד שיש למלא- הקפידו לענות על כל הסעיפים.

את התשובות המילוליות יש לכתוב באנגלית, במידה וקיימת בעיה עם האנגלית אנא צרפו תשובות מילוליות בעברית בקובץ word נפרד תוך הפניה לשאלה/סעיף המתאים בצורה ברורה.

3 נק' בונוס מובטחות לכל קבוצה שתשבי באנגלית בגוף הקובץ

את התרגיל יש להגיש לא יאוחר מה-27/12 בשעה 14:00.

בכל מקום בו יש למלא מקטע קוד תופיע ההערה:

```
#Your code here:
```

השאלות במטלה מבוססות באופן ישיר על ההדגמות שראינו ותרגלנו בכיתה - העזרו בהן לצורך חזרה והדגמה.

שימו לב שהניסוח באנגלית בגוף הטופס מבהיר מה יש לעשות והגרסה העברית נועדה להקל על הבנת המטלה

ניקוד השאלות: שאלות 1-2 10 נק', שאלות 3-6 20 נק'.

שאלות התרגיל:

1. ייבאו את הספריות והמודולים המתאימים בהם תעשו שימוש לאורך כל פתרון התרגיל.

הצעה: הכניסו תחילה את פקודות היבוא הנדרשות באופן כללי ולאחר מכן הוסיפו כל יבוא לו תזדקקו כאשר יתעורר הצורך

הקוד טוען עבורכם נתונים בשם tips לאובייקט בשם df מסוג dataframe של pandas.

הטבלה מכילה מידע על חשבונות ארוחות במסעדה והטיפים שניתנו על אותן ארוחות.

2. בעזרת הפונקציות/מתודות describe ו-shape הציגו סיכום של הנתונים. וודאו שכל העמודות מוצגות כולל הקטגוריאליות.

3. הציגו את הערכים הבאים:
   A. כמה ערכים ייחודיים יש לעמודה time?
   B. כמה מופעים יש לכל קטגוריה ייחודית (unique) מסעיף א'? (מהן שכיחויות כל הקטגוריות מסעיף א')
   C. מהו הממוצע ומהו החציון של העמודה total_bill?

4. בצעו את הפעולות הבאות:
   A. הוסיפו עמודה חדשה בשם tip_percent (אחוז הטיפ מסך כל החשבון במסעדה) ל-dataframe הקיים df, ערכי העמודה יוגדרו להיות:
   tip/total_bill*100
   B. הוסיפו עמודה חדשה בשם high_tip ל-dataframe הקיים df והגדירו את ערכיה להיות 0. לאחר מכן, כפי שהודגם בכיתה הגדירו בעמודה החדשה ערכים ל-1 אם באותה שורה בעמודה tip_percent הערך גבוה מהחציון (median).

   C. הציגו היסטוגרמה של ערכי העמודה high_tip שיצרתם בסעיף הקודם.

5. קיבוץ ערכים לפי קטגוריות (groupby):
   A. צרו אובייקט קיבוץ (groupby) בשם df_smoker באמצעות העמודה הקטגוריאלית smoker.
   B. צרו טבלה (dataframe) חדשה בשם Agg_df_smoker אשר מסכמת (באמצעות הפונקציה agg) את העמודות ["total_bill", "tip"] כך שלכל אחת משתי העמודות יוצגו הממוצע, החציון וסטיית התקן באופן הבא:
   [np.mean,np.median,np.std]
   הציגו את הטבלה
   C. ענו על השאלות המילוליות הבאות בכן או לא (מחקו את המיותר בקובץ):
   האם מעשנים נותנים טיפים גבוהים יותר בממוצע מלא מעשנים (tip)?
   האם למעשנים יש חשבון גבוה יותר בממוצע (total_bill)?

6. בסעיפים הבאים השתמשו בטבלה חדשה df_numeric המכילה משתנים מספריים בלבד (int/float) שנוצרה עבורכם טופס התרגיל.
   A. המירו את df_numeric למערך numpy בשם array_numeric באמצעות הפונקציה to_numpy.
   B. צרו מערך numpy חדש בשם array_numeric_shekel הזהה למערך array_numeric אך הכפילו את העמודות 0 ו-1 במערך החדש ב-3.16 כדי להמיר את העמודות שהיו מחירי החשבון במסעדה והטיפ (במערך החדש השמות אבדו) לשקלים.
   C. המירו את המערך array_numeric_shekel חזרה ל-dataframe של pandas בשם df_numeric_shekel ותנו שמות לעמודות של הטבלה החדשה שיהיו זהים לשמות של הטבלה המקורית df_numeric למעט השינויים הבאים:

```
"total_bill" --> "total_bill_shekels"
"tip" --> "tip_shekels"
```

בהצלחה!!