מבוא לניתוח נתונים – תרגיל מספר 5 – הכנה למבחן

תרגיל זה מדמה מבחן, מבחינת סוג השאלות ורמת הקושי.

- תאריך הגשה: 8/6/2021
- ההגשה לבד או בזוגות. במקרה של הגשה בזוגות -רק אחד מבני הזוג מגיש, אך יש להקפיד לרשום את השם ות.ז. של שני הסטודנטים שהגישו. מי שלא רשום על אף עבודה לא יקבל ציון, ולא יוכל להצטרף לעבודה בדיעבד.
- במידה וסעיף מסוים לא רץ – הציון על סעיף זה הינו אפס.
- יש להראות את הקוד דרכו הגעתם לפתרון. סעיף עם תשובה נכונה אך ללא קוד יקבל ציון אפס.
- התרגיל הינו אישי וניתן לבצעו בזוגות בלבד. אין להעביר פתרונות או חלקי פתרונות בין סטודנטים בכיתה. העברת התרגיל על כל המשתמע מכך לאנשים אחרים שאינם רשומים לקורס אסורה. כל הנ"ל הינם עבירות משמעת אשר יועברו לבחינת ועדת המשמעת של האוניברסיטה.
- הקפידו להגיש עבודה ברורה עם הסברים במידת הצורך. אל תשאירו אותנו במתח! כתבו באמצעות markdown על איזה סעיף אתם עונים. על סעיף ללא הסברים (או במידה ולא ברור באיזה סעיף מדובר) יורדו לפחות חצי מהנקודות.
- יש להגיש קובץ אחד בלבד, מסוג ipynb.
- איחורים בהגשה מכל סיבה שאינה נמצאת בתקנון הרשמי של האוניברסיטה: על כל יום איחור בהגשה יורדות 5 נקודות בציון התרגיל באופן אוטומטי.
- יש להשתמש בכלים שנלמדו במהלך הקורס. אין להשתמש בלולאות for

בעבודה זו תנתחו נתוני מכירות של אבוקדו.

יש להגיש את התשובות בתוך הקובץ:

Hw5_solution_template

הסבר על הדאטה:

- `Date` - The date of the observation
- `AveragePrice` - the average price of a single avocado
- `type` - conventional or organic
- `year` - the year
- `Region` - the city or region of the observation
- `Total_sold` - Total number of avocados sold
- Small_sold - Total number of small avocados sold
- `Large_sold` - Total number of large avocados sold
- XL_sold - Total number of XL avocados sold
- Total_bags – total number of avocado bags sold

1. (15%) הציגו טבלה שמציגה <u>כמה</u> אבוקדו נמכר בכל איזור בכל שנה.

2. (10%) צרו dataframe חדש שמכיל רק איזור (region) שמתחיל באות הראשונה של שמכם, ואת השנה (year) 2016. במידה ואין איזור באות הראשונה של השם שלכם, ביחרו את האיזור עם האות הקרובה ביותר (לא משנה לאיזה כיוון באלף-בית). כמה שורות יש ב dataframe שלכם?

3. (10%) על אותו ה dataframe שיצרתם בשאלה 2: מהו הממוצע מכירות?

4. (15%) על אותו ה dataframe שיצרתם בשאלה 2: מהי סך הכל כמות שקיות האבוקדו שנמכרו (העמודה total bags) בחודש ספטמבר (חודש מספר 9, בעמודה date) ?

<u>השאלות הבאות מתייחסות לכל הדאטה:</u>

5. (15%) צרו ויזואליזציה שמראה השוואה בין כמות האבוקדו (העמודה total_sold) משני הסוגים (העמודה types)

6. (20%) הציגו את הקורלציה עבור העמודות averagePrice, total_bags, total_sold, type . שימו לב שיש לטפל תחילה בעמודה type כי היא אינה מספרית.

7. (15%) צרו scatterplot עבור הקורלציה הכי חזקה שמצאתם עם צבע (hue) על פי השנה. במידה ויש תיקו קחו את אחת מהקורלציות הכי חזקות.

דוגמאות לשאלות נוספות בסגנון המבחן (אין צורך לפתור עבור תרגיל 5):

- במידה ויש ערכים חסרים באחת העמודות, השלימו אותם בעזרת חציון/ממוצע/שיטה אחרת מתאימה
- הראו את התפלגות הנתונים עבור total_sold הקפידו שציר y יהיה ב Logscale (המשמעות: צרו הסטוגרמה).
- צרו טבלה שמראה את המחיר הממוצע (העמודה averagePrice) לכל סוג אבוקדו (העמודה type) על פני כל השנים (העמודה year).
- צרו מודל שמנבא את total_sold לפי: region, type, averagePrice, total_bags. מהי רמת הדיוק של המודל שיצרתם?