Python Libraries - Moed A Exam

Question 1

הכנת מידע

לפניך קבצי המידע הבאים:

Customers.csv, Items.csv, order_item.csv, orders.csv

חקור את המידע

- הדפס לכל טבלה (בפעולה אחת פר טבלה) את פרטי שמות העמודות , הטיפוס, וכמות העמודות שאינן Null (ועוד)
- הדפס לכל טבלה (בפעולה אחת פר טבלה) את שמות העמודות ,הערכים הממוצעים, ערך MIN, ערך MAX חציון (ועוד)
- הדפס לכל טבלה את 5 הערכים הראשונים

מחיקת כפילויות

בטבלת items קיימים מוצרים שמופיעים מספר פעמים (מוצר עם אותו השם) והוחלט לשמור במוצרים בעלי אותו שם, את המוצר עם הכי הרבה פרטים not-null, ואת היתר למחוק.

רמז: (עקוב אחרי הדוגמא שעשינו בשיעור ויישם אותה כאן),

דרך פעולה-

- הוסף עמודה חדשה הסופרת כמה ערכי not-null יש בכל שורה
- מיין לפי שם המוצר + העמודה החדשה שייצרת, מהגדול לקטן
- הסר את הכפילויות לפי שם המוצר (שמור רק את המופע הראשון)

עדכון טבלת order_item:

- החלף את ה- id של המוצרים עם הכפילויות ל- id של המוצר עם הכי הרבה פרטים not-null. לדוגמא:

48	Camera	Photography Equipment	2039.08	471	Panasonic	TRUE
28	Camera		695.71		Sony	FALSE

כאן, מכיוון שתוסר המצלמה עם id 28 יש לשנות בטבלת order_item את כל המכירות של מוצר המצלמה עם id 28 ל- 48

מחק מטבלת customers את כל הלקוחות אשר חסר להם 3 או יותר שדות

מחק מטבלת order_item את כל המכירות של הלקוחות שהסרת בסעיף הקודם

Question 2

ניתוח מידע

השתמש בקבצי המידע הבאים:

Customers.csv, Items.csv, order_item.csv, orders.csv

1. מה מחיר ממוצע של פריט
2. מי הלקוח שרכש הכי הרבה מוצרים
3. הוסף עמודה total_price לטבלת order_item, המחשבת את עלות הקנייה הכוללת (מחיר מוצר * כמות)
4. הצג כל קנייה ואת המחיר הכולל שלה
5. מצא מה היא הקנייה הכי יקרה שבוצעה? הכי זולה? ממוצע?
6. מטבלת customers, צור טבלת PIVOT ובה הצג כל לאום כאינדקס, ואת ה- gender כעמודות.
   הצג כמה לקוחות משתייכים לאותו המגדר באותו הלאום.
   לדוגמא (פלט חלקי של הטבלה...)-

gender	Agender	Bigender	Female	Genderfluid	Genderqueer
nationallity					
Argentina	1	0	10	1	0
Brazil	0	0	6	0	0

<u>חקור והצג **בגרף** את הפרטים הבאים:</u>

המלצה: הסר שורות עם ערכים ריקים

1) פילוג לקוחות לפי מגדר , gender
2) פילוג כמות לקוחות לפי לאום
3) הסטוגרמת פילוג לקוחות לפי גיל
4) גרף המציג כמות לקוחות חדשים שהצטרפו, לפי שנים
5) גרף המציג כמות מכירות לפי חודשים
6) הסטוגרמת פילוג כמות קניות לפי גיל (יש להחשיב כל קניה בספירה אחת. אין צורך לספור כמות מוצרים וכו') – האם יש מגמה ליותר/פחות קניה בטווחי גיל מסויימים?
7) *אתגר: עלות קניות לפי לאום. כלומר- נסכום את סך עלות הקניות שנעשו לכל המוצרים עבור כל לקוח, נחבר ביחד את כל קניית הלקוחות מאותו לאום – ואז נראה כמה עלות יש פר לאום

Question 3

❓ שאלות:

1. `df.dropna`

א. מוחקת כפילויות מתוך עמודות

ב. מוחקת שורות שמכילות ערכים חסרים (NaN)

ג. מחליפה ערכים חסרים בערך ברירת מחדל

ד. מוחקת את עמודת האינדקס מהטבלה

---

2. `df['city'].nunique`

א. מחזיר את שם העיר שהופיעה הכי הרבה פעמים

ב. מחזיר רשימה של כל הערים בטבלה

ג. מחזיר את מספר הערכים השונים בעמודה 'city'

ד. מוחק ערכים כפולים מהעמודה 'city'

---

3. `df['gender'].replace({'M': 'Male', 'F': 'Female'})`

א. ממיר את הערכים למספרים לפי סדר הופעה

ב. מחליף את ערכי העמודה gender לפי מפת המרה

ג. מסנן את הרשומות שבהן gender הוא M או F

ד. מחבר בין ערכים שונים בעמודה gender

---

4. `df.sample(5)`

א. בוחר את 5 הערכים הקטנים ביותר בטבלה

ב. מחזיר את כל הערכים של עמודת sample

ג. בוחר 5 שורות אקראיות מהטבלה

ד. יוצר עותק של הטבלה עם 5 עמודות בלבד

---

5. `df['score'].fillna(0)`

א. מוחק את כל הערכים בעמודה score

ב. מסמן את השורות בהן הערך הוא אפס

ג. ממלא את הערכים החסרים בעמודה score עם אפס

ד. ממיר את כל הערכים באפסים

6. `df.drop_duplicates`

א. מוחקת שורות שמכילות ערכים חסרים

ב. מחליפה ערכים כפולים בערך ממוצע

ג. מוחקת שורות כפולות בטבלה

ד. מוחקת את כל הערכים הייחודיים בטבלה

---

7. `df['price'].nlargest(3)`

א. מחזיר את 3 הערכים הקטנים ביותר בעמודת price

ב. מחזיר את 3 הערכים החסרים בעמודת price

ג. מחזיר את 3 הערכים הנפוצים ביותר בעמודת price

ד. מחזיר את 3 הערכים הגדולים ביותר בעמודת price

---

8. `df['city'].value_counts`

א. מציג את הערכים המופיעים בעמודה בסדר אקראי

ב. מחזיר את מספר ההופעות של כל ערך בעמודה

ג. מוחק כפילויות בעמודת city

ד. ממיר ערכים טקסטואליים למספרים

---

9. `df['quantity'] > 100`

א. בודק אילו שורות בעמודה quantity שוות בדיוק ל־100

ב. מסנן רק את השורות שהכמות בהן שווה לאפס

ג. יוצר סדרה בוליאנית של True/False לפי תנאי

ד. משנה את הערכים בעמודה ל־True

---

10. `()df.groupby('category').mean`

א. מחלק את הטבלה לפי קטגוריה ומחשב ממוצע לעמודות מספריות

ב. מוחק ערכים כפולים בעמודת קטגוריה

ג. מחזיר רק את הקטגוריה שהופיעה הכי הרבה

ד. משכפל את הקטגוריות לפי מספר ההופעות שלהן

הנחיות הגשה:

יש לייצר **2 ריפו ב- GITHUB** :

- אחת לפתרונות ההגשה של **היום**
  - כאן לא יהיה COMMIT <u>אחרי התאריך של היום</u>
- שנייה ליתר הפתרונות
  - כאן לא יהיה COMMIT אחרי שישי בצהריים

לטובת התשובות לשאלות האמריקאיות – לייצר קובץ טקסט ולציין בו מספר שאלה + מס' תשובה כדאי בנוסף לציין את טקסט התשובה גם במילים, למנוע מצב של בלבול,

לדוגמא:

שאלה 18- תשובה ב' "מחיקת הכפילויות"

יש לשלוח את הלינק לריפו למייל-

pythonai250824+EXAMAILIB@gmail.com

נא לציין במייל:

- שם פרטי
- שם משפחה
- ימי הקורס

בהצלחה!