**2016**

1. **הוצאת הנתונים:**

קובץ הנתונים-

IDF\_ABM\_16.2.15\_wide.csv

הקובץ המקורי –

שאלונים, יש רק זמן 1

**שאלון PCL**

C:\Users\nogag\Documents\birocracy\PTSDClassifier\PTSD\Data\2016\questionnaire5\_PCL.csv

**שאלון דיכאון**

C:\Users\nogag\Documents\birocracy\PTSDClassifier\PTSD\Data\2016\questionnaire4\_PHQ9.csv

**קבצים מקוריים:**

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה נובמבר 2013\השלמות 5.5.13\questionnaire 4\_PHQ-9\_5.5.sav

D:\PTSD\אילן, מחקר צבא, 2012-2015\DATA\שאלונים\השלמות מניעה 20.1.2013\questionnaire 4\_PHQ-9.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה דצמבר 2012\questionnaire 4\_PHQ-9.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה נובמבר 2013\questionnaire 4\_PHQ-9.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה נובמבר 2013\questionnaire 5\_PCL.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה דצמבר 2012\questionnaire 5\_PCL.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\השלמות מניעה 20.1.2013\questionnaire 5\_PCL.sav

אילן, מחקר צבא, 2012-2015\DATA\שאלונים\מחקר מניעה נובמבר 2013\השלמות 5.5.13\questionnaire 5\_PCL\_5.5.sav

**קובץ ההמרה ל-SPSS**

[spss 2 csv-2016.ipynb](http://localhost:8888/notebooks/spss%202%20csv-2016.ipynb)

1. **קובץ שילוב הנתונים**
2. **רשימת הפיצ'רים המקוריים:**

**אקספלורציה**

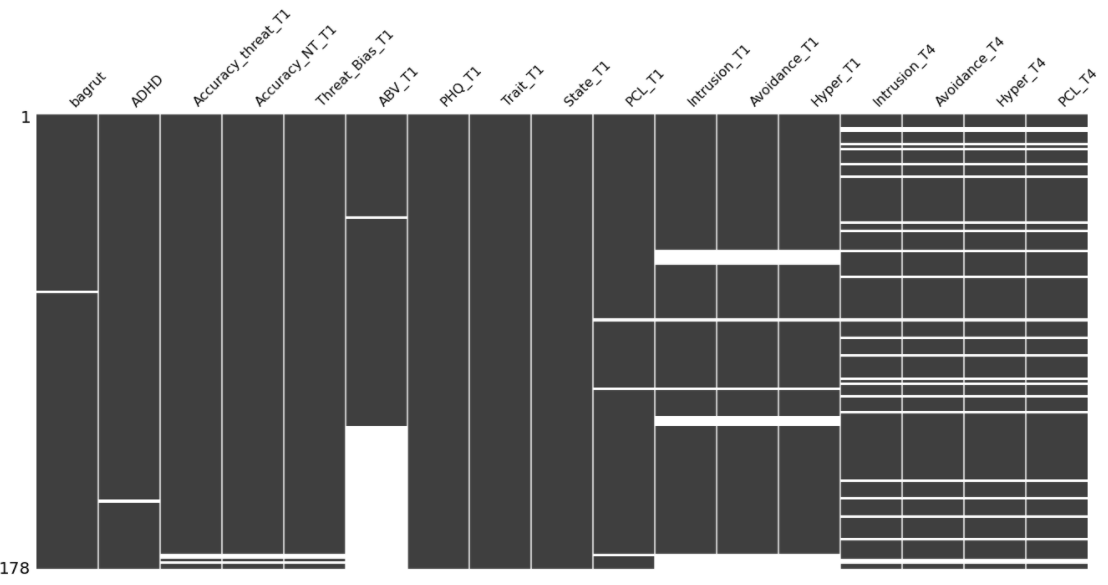
**נתונים חסרים**

נראה את התפלגות הנתונים החסרים על תת-הסט של הפיצ'רים בלי פירוט ברמת השאלה הבודדת:

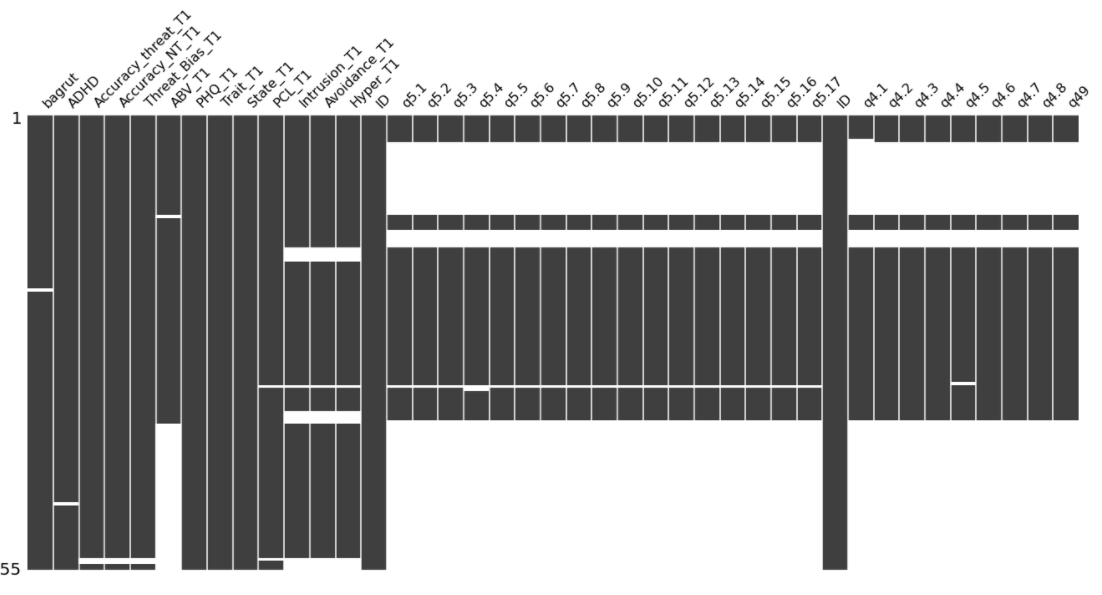
נחלק את החוסרים לארבעה סוגים:

1. נבדקים עם חוסרים ב-target
2. נבדקים עם חוסרים ב-ABV
3. נבדקים עם חוסרים בכל הנתונים הקוגנטיביים
4. נבדקים עם חוסרים בנתוני ה-PCL
5. נבדקים עם חוסרים רנדומליים

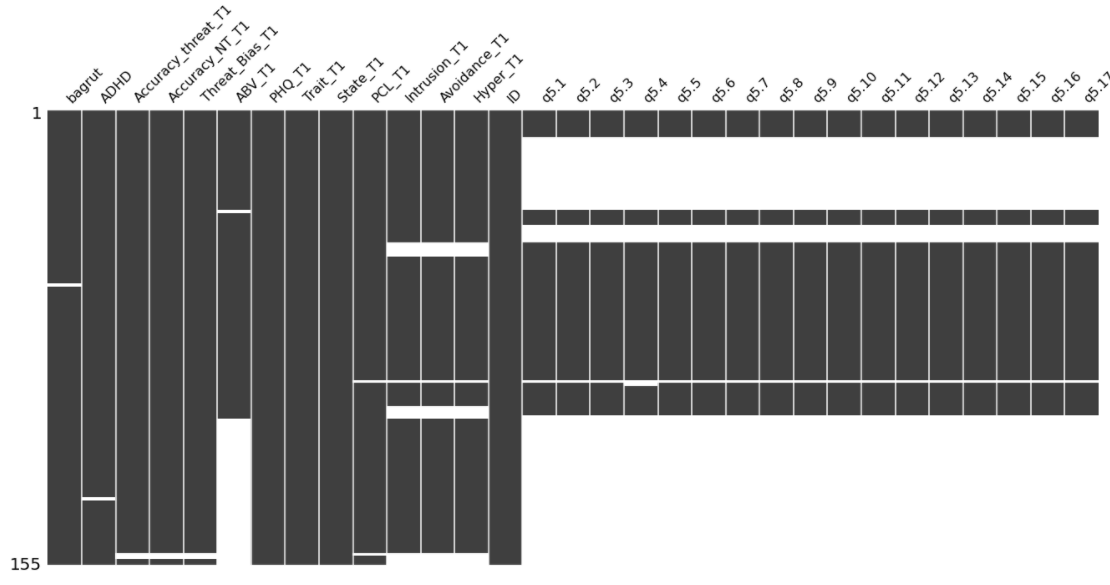
נוציא נבדקים עם חוסרים ב-target



ניתוחי המשך של חוסרים –

יש כמות גבוהה של חוסרים בשאלונים, נראה שהחוסרים האלה מקבילים לחוסרים ב-ABV   


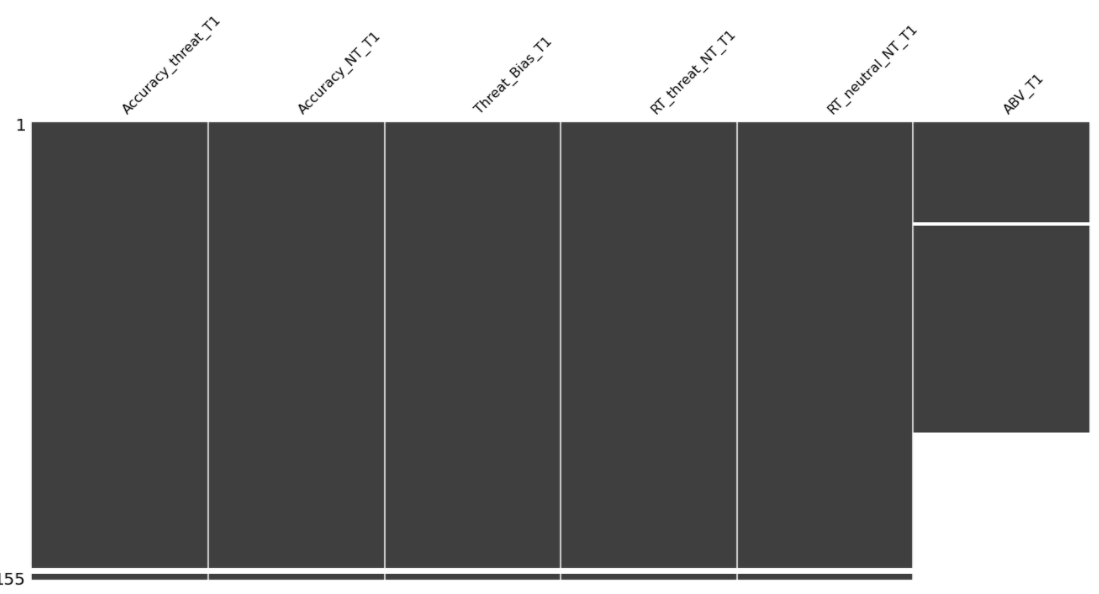
**חוסרים ב-PCL1:**

נראה שיש מקרים שאפשר למלא את החוסרים בדאטה-פריים המקורי ע"י סכימה של השאלונים.  


**חוסרים ב-PHQ9** – די משעממים

**חוסרים בנתונים קוגנטיביים:**

מלבד ה-ABV, אין הרבה חוסרים, מה שיש נראה מקביל לכולם.



**לשאול את ירון ויאיר:**

1. **יש בעיה להשתמש בשאלונים בתור שיקלול לערכי הטראומה החסרים? -אין בעיה**
2. **האם להשתמש ב-ABV? – לנסות להבין מה הבעיה, לחשב מחדש**
3. **האם להשתמש בשאלונים? – לבדוק אם זה עוזר**
4. **האם המיפוי שיש לי הוא נכון? (מבחינת שמות המשתנים) – כן**
5. **האם להשמיט נתונים בלי מדדים קוגנטיביים? לא, אפשר לנסות אח"כ אם זה עוזר**

**ניקוי נתונים**

**הסרת שורות כפולות** באמצעות הפקודה df.drop\_duplicates("ID")

נראה שאין שורות כפולות לחלוטין, אבל יש מספר שורות לאותו ID בשאלונים. נראה שהשורות עם ה-ID הכפול כוללות שורה עם חוסרים בכל הערכים מלבד ה-ID.  
כל זה נכון מלבד נבדק – 2192 שיש לו ממש שני ערכים שונים... ונראה שבזמן 4 יש לו פחות טראומה מאשר בזמן 1, בנוסף הוא מופיע (4!) פעמים בדאטה פריים הסופי. O\_O

**נסיר אותו בינתיים**

**לקחת לפי הקובץ הסופי**

לאחר הטיפול בכפילויות ה-ID, אין יותר ערכים כפולים.

**קריטריונים להסרת ערך של משתנה:**

**לשאול את ירון ויאיר: האם להסיר נתונים קוגנטיביים של מי שיש לו פחות מ-70% דיוק על ה-DP.**

**האם להסיר עוד משהו?**

**לא להסיר**

**קריטריונים להסרת נבדק שלם:**

נסיר נבדקים שאינם בתנאי הביקורת (נופלים 615)

נסיר נבדקים שיש להם חוסרים ב-target. (נופלים 23)

**נתונים מלוכלכים:**

נראה שמלבד ערכים חסרים, מדדי ה-PCL ושלושת תת-הסטים שלו מתאימים בין הקובץ המלא לבין קובץ השאלות.

**עיבוד הנתונים:**

1. דמוגרפיים:

* 'bagrut': האם יש לנבדק בגרות, משתנה בינארי
* ADHD: האם הנבדק סובל מהפרעת קשב וריכוז, משתנה בינארי

**תהליך העיבוד**: בודקים האם הערך שווה ל-'yes', והופכים את התוצאה ל- int

**סוג משתנה**: בינארי, int

1. שאלונים

* 'PHQ\_T1'
* Trait\_T1'
* State\_T1

**תהליך העיבוד**: אין  
**סוג משתנה**: משתנה סדר/ רווח דיסקרטי

1. שאלון טראומה

* PCL\_T1
* Intrusion\_T1
* Avoidance\_T1
* Hyper\_T1
* **סוג משתנה**: משתנה סדר/ רווח דיסקרטי

**עיבוד המשתנים**: אין

**הערות**:

1. כל המשתנים אינם רלוונטיים ל-2018 בלי התאמה
2. גם עם התאמה, ההימנעות (חלקית) וההיפרסנסטיביות לא רלוונטיים ל 2018
3. אפשר לחפש את המשתנה הבינארי של הטראומה
4. נתונים קוגנטיביים

* Accuracy\_threat\_T1'
* Accuracy\_NT\_T1'
* 'Threat\_Bias\_T1

**תהליך העיבוד**

**סוג משתנה**: משתנה רציף

1. פיצ'רים של אימון

* control
* placebo
* train\_4

**תהליך העיבוד**: pd.concat((df\_2016,pd.get\_dummies(df\_2016.Group)),1)

**סוג משתנה**: משתנה בינארי מומר ל-int

1. משתני מטרה

**משתני מולטיטאסק:**

* Intrusion\_T4
* Avoidance\_T4
* Hyper\_T4

**תהליך העיבוד**: אין

**סוג משתנה**: משתנה סדר/רווח דיסקרטי

**משתנים מספריים**

משתנה PCL\_T4, ציון ה-PCL בזמן 4

**סוג משתנה**: משתנה סדר דיסקרטי

**משתנים בינאריים**

משתנה ה- target\_feature שאנחנו מגדירים כאן:

האם PCL\_T4 גדול ממש מ-39, מומר ל-int

**סוג משתנה**: משתנה בינארי מומר ל- int

משתני מטרה באקסל

'CAPS\_Intrusion\_T4\_CLI',

'PTSD\_CAPS',

'PTSD\_CAPS\_FUNC',

'CAPS\_TOTAL\_17',

'PCL\_T4',

'Intrusion\_T4',

'CAPS40',

'CAPS30',

'PCL4\_40',

'PCL4\_30',

משתני מטרה נוספים