Exploratory Data Analysis

פרטי הנתונים:

הנתונים מחולקים לשתי קבצים. בכל קובץ רשימה של רצפים, ולכל קובץ מספר תכונות. בקובץ א' נמצאים רצפי ה- DNA שממופים ל שרשרת אלפא, ובקובץ השני ישנם רצפי DNA שממופים לשרשרת בטא.

גודל הנתונים:

TRA:

455,564 שורות (רצפים) ו- 31 עמודות (תכונות)

TRB:

1,420,003 שורות (רצפים) ו- 31 עמודות (תכונות)

שמות העמודות:

['GlutamicAcid', 'Glycine', 'Serine', 'Isoleucine', 'Cysteine',

'Methionine', 'NegativelyCharged', 'Aliphatic', 'Tyrosine',

'FrameShift', 'Valine',

'PositivelyCharged', 'Asparagine', 'Lysine',

'Polar', 'StopCodon', 'MolecularMass', 'Arginine',

'length', 'Glutamine', 'Aromatic',

'IsoelectricPoint', 'Histidine',

'Proline', 'Tryptophan', 'AsparticAcid', 'Phenylalanine', 'Leucine',

'Threonine', 'Alanine', 'Hydrophobicity']

הסבר העמודות:

כל עמודה היא תכונה של רצף מסוים.

- ישנם 20 שמות של חומצות אמינו, כאשר הערך של אותה תכונה מייצג את אחוז החומצה האמינית מכלל רצף החלבון, שמתורגם מרצף ה- DNA.
- ישנם 8 תכונות ביוכימיות של הפפטיד:

  'NegativelyCharged', , 'Aliphatic' ,'PositivelyCharged', 'Polar', 'MolecularMass', 'IsoelectricPoint', 'Hydrophobicity', 'Aromatic'

- ישנם 3 תכונות של הרצף עצמו:

  'FrameShift', 'StopCodon', 'length'

כפילויות:

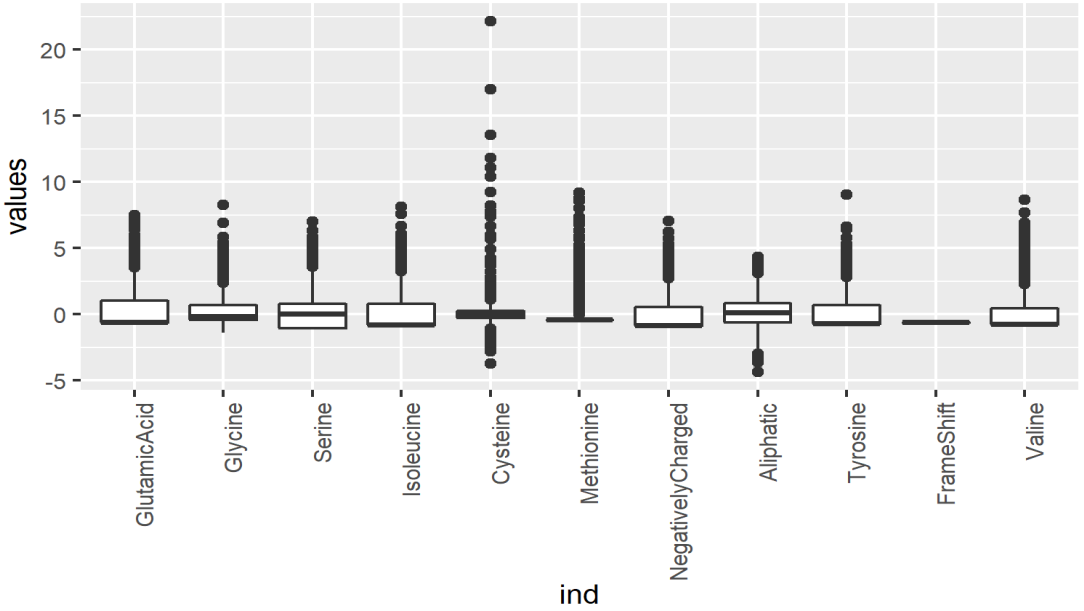
לא נמצאו רצפים כפולים

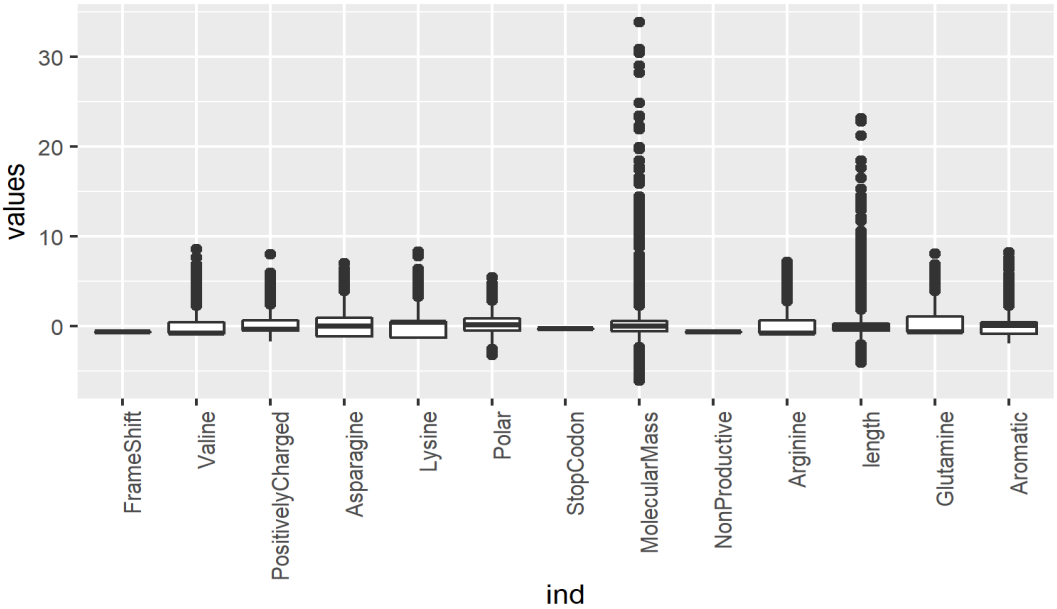
נתונים חסרים:

לא נמצאו נתונים חסרים

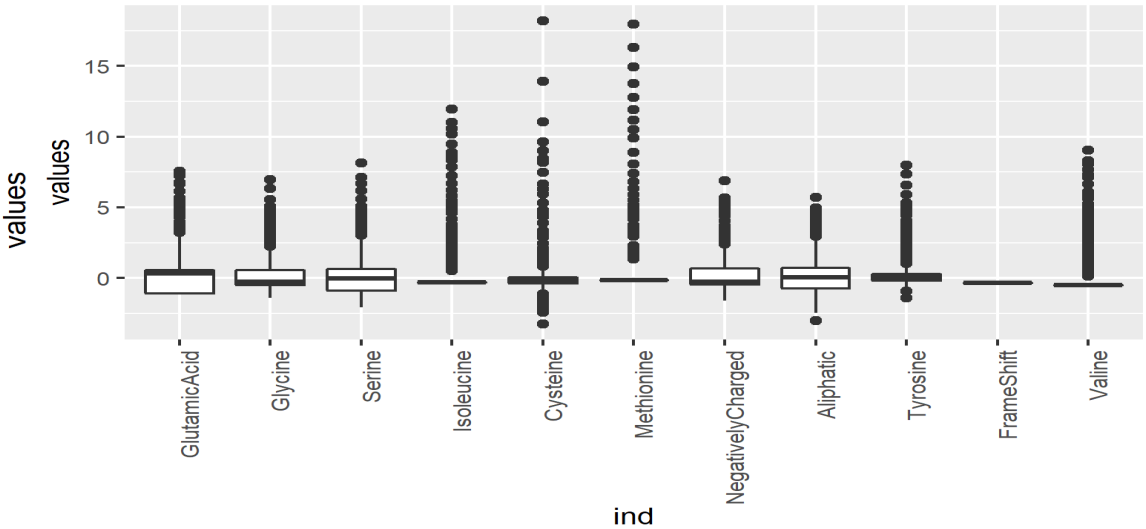
סטטיסטיקות:

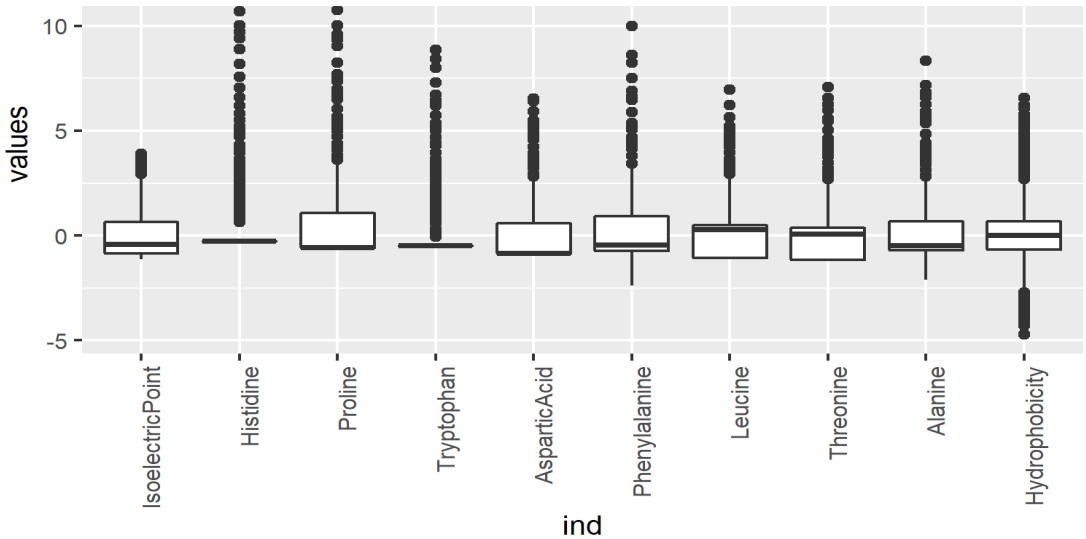
boxplots TRA:

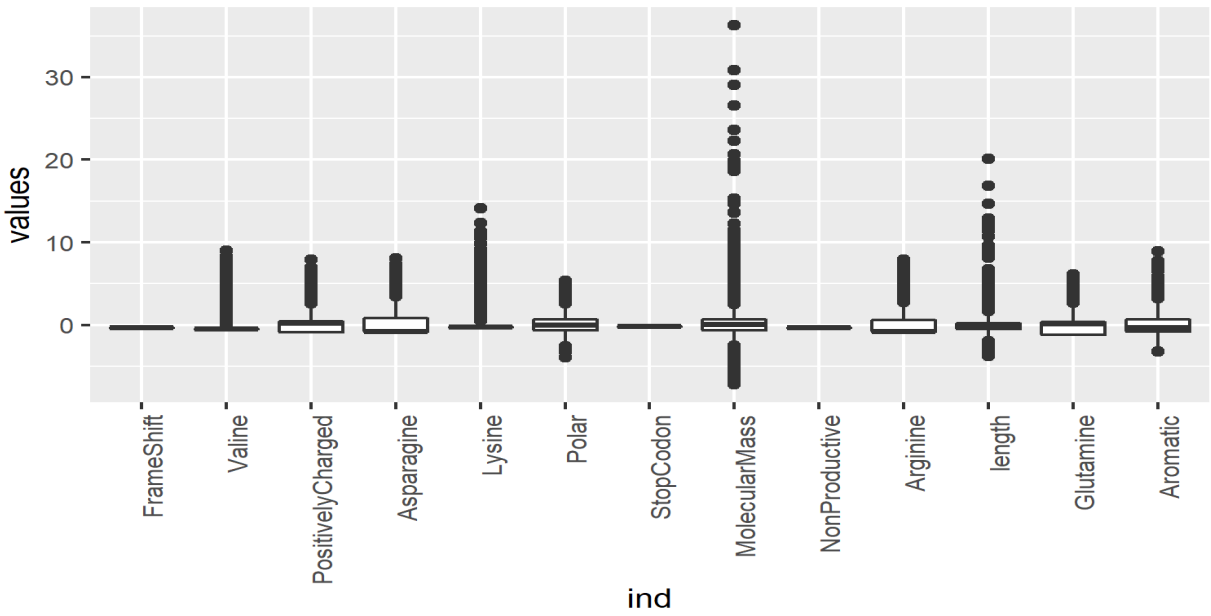



values
30
20
10
0
FrameShift
Valine
PositivelyCharged
Asparagine
Lysine
Polar
StopCodon
MolecularMass
NonProductive
Arginine
length
Glutamine
Aromatic
ind

:TRB


values
15
10
5
0
GlutamicAcid
Glycine
Serine
Isoleucine
Cysteine
Methionine
NegativelyCharged
Aliphatic
Tyrosine
FrameShift
Valine
ind


values
10
5
0
-5
IsoelectricPoint
Histidine
Proline
Tryptophan
AsparticAcid
Phenylalanine
Leucine
Threonine
Alanine
Hydrophobicity
ind



<u>outliers:</u>

1. ניתן לראות שבפיצרים מסוימים ישנם כמות גדולה של outliers, למשל בציטוזין וב length.
2. ניתן לראות שיש שוני במספר ה- outliers בין tra ל-trb . למשל, למתיונין יש יותר outliers ב- trb מאשר ב- tra.

למסקנה החלטנו להשאיר את הנתונים כמו שהם, ולא לבצע בהם ניקוי מ outliers, מכיון ש:

א. הניקוי היא פעולה מסובכת וארוכה
ב. לא בטוח שה- outliers האלה נובעים מטעות. יכול להיות שדווקא אם נוריד אותם, נוריד את איכות הנתונים.

<u>קורלציה בין משתנים:</u>

מצורפת טבלה של חישוב קורלציה בין המשתנים לפי מתאם פירסון.

- ניתן לראות שלח"א מסוימות, יש קורלציה עם התכונות שלהם. למשל, אחוז חומצה גלוטמית שהיא פולרית וטעונה שלילית, נמצא בקורלציה גבוהה יחסית עם התכונות "Polar" ו- "NegativelyCharged".
- לשאר התכונות שאין להם קשר ישיר עם התכונות של ח"א, שהם: "FrameShift", "StopCodon" ו "length", אין קורלציה עם אחת מח"א, חוץ מקורלציה הפוכה מסוימת שנמצאה בין "Cysteine" ל "length".

לסיכום:

איכות הנתונים נראית טובה, חוץ מ outliers שהחלטנו להתעלם מהם בשלב זה. ראינו קורלציה טובה בין פיצ'רים שציפינו למצוא בהם קורלציה (ח. גלוטמית ו polar) וגם קורלציה שלא ציפינו לה (בין "Cysteine" ל "length"). ניתן להמשיך לחקור קורלציות כאלה, וכן קורלציות אחרות שלא דנו בהם כאן.