עיבוד מקדים:

השלבים שכבר קיימים –

1. להעיף נבדקים בלי אינדקס מטרה
2. להעיף נבדקים שחסר להם זמן 1 או זמן 2
3. את שאר הפיצ'רים החסרים להשלים עם מייס
4. להעיף נבדקים עם כמות גבוהה של אאוטליירים

שיפורים:

* להוסיף פיצ'רי אשכנזיות ומזרחיות, לשקול מחדש להכליל את הנתון הגנטי עם הרבה ערכים חסרים
* לבדוק לעומק יותר את ההשלמה של החסרים:

1. האם צריך להפריד לפיצ'רים בדידים ורציפים?
2. האם יש לנו פיצ'רים קטגוריאליים?
3. האם מייס בכלל מתאים לבדידים?

* האם יש לנו ערכים לא תקינים?
* האם להחזיר פיצ'רים אאוטליירים למקום שלהם? או לסמן כחסרים ולהשלים? או לא לגעת?
* מה הפיצ'רים החשובים?
* האם יש לנו פיצ'רים שכדאי לפצל לכמה?
* לבדוק תלות בין ערכים חסרים לבין ערך המטרה
* ליצור משתנה שמצביע על ערכים חסרים?

מטלות ספציפיות:

* לעשות פלוט של סכום החסרים
* לעשות פלוט של סכום האאוטליירים
* לעשות קורלציות של מטרה ושל הערכים החסרים של כל אחד מהפיצ'רים
* לעשות קורלציות של מטרה ושל האאוטלייריות של כל אחד מהפיצ'רים
* לעשות חשיבות פיצ'רים
* התבוננות בהתפלגות לפניואחרי אימפוטציה, ולפני ואחרי פלורינג