<u>פרויקט סופי – שיטות במדעי הנתונים</u>

הגשה 04/07/2022

בפרויקט זה תעבדו על נתונים אמיתיים אנושיים של רמות ביטוי של גנים.
אורך העבודה לא יעלה על 6 עמודים. (אפשר לשים חומר נוסף בנספחים. החלק שיבדק הם 6 העמודים).
על העבודה לכלול הסבר על שיטות, תוצאות ומסקנות מהאנליזה.
עליכם להסביר בדיוק באיזה שיטות השתמשתם, כיצד בוצע כל חישוב, ובאיזה פונקציה עיקרית השתמשתם (למשל lm). יש להציג תרשימים רלוונטיים ותוצאות רלוונטיות ולהסביר.
עליכם להגיש את העבודה הכתובה ובנוסף קובץ עם הקוד שלכם כך שהוא מתועד היטב (יש להוסיף הסברים לקוד).

עליכם לבחור איבר אחד מבין (אפשר איבר אחר באישור מרצה):
"Adipose - Subcutaneous", "Liver", "Thyroid", "Muscle - Skeletal", "Nerve - Tibial", "Skin - Not Sun Exposed (Suprapubic)", "Whole Blood", "Lung", "Esophagus - Muscularis"
*שימו לב שבאיבר שבחרתם יש מספיק דגימות (מעל 50).

עליכם לבצע על האיבר שבחרתם עיבוד נתונים לפי המפורט למטה.

הפרויקט מורכב משני חלקים.

1. **שלב עיבוד ראשוני של הנתונים**: עליכם להסביר בקצרה ולהראות גרפים רלוונטיים לגבי מציאת אאוטליירים, נירמול ופילטור הנתונים עם הסבר קצר של האיבר שעליו בחרתם לעבוד.

2. **תיקון הטיות – זיהוי פקטורים (תכונות) שמוסיפות "רעשים" לנתונים**: בשלב זה עליכם לחקור את הנתונים והשפעות עליהם. עליכם להחליט אילו תכונות מהוות "רעש" ומקלקלות את ערכי הנתונים ע"י שימוש בשיטות שנלמדו בקורס. ראו למטה הסבר על תכונות רלוונטיות.
עליכם להראות גראפית וחישובית האם מצאתם השפעה של כל תכונה על רמת הביטוי של הגנים. בנוסף יש לחקור קשרים בין התכונות.
עליכם להיות יצירתיים ולהשתמש בכמה שיטות שלמדתם בקורס כדי להראות אילו תכונות משפיעות על רמת הביטוי של הגנים (הערך של הגנים), ביניהם - PCA, למידת מכונה, קורלציות, רגרסיה לינארית וכו.
עליכם לבנות מודל למידת מכונה לזיהוי רמת הביטוי של הגנים (הפיצ'רים) <u>לפחות לסוג אחד של תכונה</u> של הנתונים ("רעש"). יש להסיק מתוצאות הדיוק של המודל אם התכונה משפיעה על הנתונים.
הנחיה:
PCA יש לבצע לכל חמשת התכונות באופן הבא – יש לייצג בצביעה בגרף ה PCA את התכונה (למשל לצבוע דגימות לפי הגיל מצעיר למבוגר או הזמן שעבר מרגע המוות עד לקיחת הדגימה).

עליכם להסביר אילו תכונות משפיעות ופוגמות בנתונים ולהחליט במסקנות הפרויקט אילו תכונות יהיה צורך לתקן בדאטה.

למטה יש הסבר על תכונות שיכולות להשפיע על רמות הביטוי של הגנים. עליכם לבדוק האם יש השפעה לכל תכונה על רמות הביטוי של הגנים באיבר שאותו בחרתם.

<u>להלן הסבר על הנתונים וקוד עזר:</u>

Description of the data:

Gene expression data
R object: Dataframe name
mat.f.coding.RData: mat.f.coding

rows – represent the genes
columns – represent the samples (the individuals)

Samples Characteristics:
R object: Dataframe name
pheno.f.RData: pheno.f

Gene names
R object: Dataframe name
gene.f.RData: gene.f
Name – the formal number of the gene
Description – the name of the gene (called Gene Symbol)

Sample Characteristics:

- SMRIN - The RNA Integrity Number, a measure of the quality of the RNA
- SMTSISCH - Interval in minutes between actual death and final tissue stabilization, NOTE: some blood samples are collected prior to the donor's death, this is indicated by negative ischemic times
- SMGEBTCH - Expression Batch ID, Batch when DNA/RNA from a sample was analyzed (the number of the experiment)
- AGE – the age of the individual in years. In 10 years intervals.
- DTHHRDY - Death Circumstances. Death classification based on the 4-point Hardy Scale:
  0 = Ventilator Case All cases on a ventilator immediately before death.
  1 = Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min.
  2 = Fast death of natural causes. Sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr.
  3 = Intermediate death. Death after a terminal phase of 1 to 24 hrs (not classifiable as 2 or 4); patients who were ill but death was unexpected.
  4 = Slow death Death after a long illness, with a terminal phase longer than 1 day (commonly cancer or chronic pulmonary disease); deaths that are not unexpected.

השתמשו בפונקציית העזר הבאה:
הפונקציה מקבלת את מטריצת pheno.f, וקטור של מספרי הדגימות ומחזירה את מטריצת התכונות של דגימות אלו. מתוך מטריצת התכונות תוכלו לבחור על אילו תכונות תירצו לעבוד ולכלול באנליזת התיקון שלכם.

```
get.pheno.mat.from.samplIDs <- function(pheno.mat, samples.vec){
  indexes      = which(pheno.f$SAMPID %in% samples.vec) #these are the column indexes of the samples
  #check the case of more than one sample for a subject
  tissue.pheno = pheno.f[indexes, ]
#second parameter is the order that we want
  ordering     = match(tissue.pheno$SAMPID, samples.vec)
  tissue.pheno = tissue.pheno[ordering,]
  if(!identical(as.character(tissue.pheno$SAMPID), samples.vec)) {print("ERROR 2: samples in norm.mat and pheno do not match")}
#this is the related pheno tabel ordered accoring to the intial subject ids order, as in #the expression table
```

```
  return(tissue.pheno)
}
```

קוד עזר:

```
samples.ids      = colnames(tissue.edata) #the ids of the subjects in this tissue
   tmp.pheno        = get.pheno.mat.from.samplIDs(pheno.f, samples.ids)
```