You will need to write code to make plots and answer the quiz questions, but you do not need to upload any of this code. Please make sure your submission still meets all project formatting requirements if you adjust anything during the quiz. When we ask for a brief explanation, please limit to 1-2 sentences. For plots, remember to label all axes. You can create the plot using matplotlib, base R, ggplot2, excel, or any other software of choice, but please do not upload hand-drawn plots.

Part 1: Dataset information

We obtained the expression data for this project from GEO. In this part of the quiz, we want you to view the dataset page in GEO, link below, and answer the following questions.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25628

Q1. Click on any of the samples listed and find the header descriptions for the expression data table. From the description, what do the expression values represent?
A1. RMA log2 intensities

Q2. Navigate to the publication associated with this dataset. Which gene do the authors identify as potentially playing a role in disease development that had never before been associated with endometriosis?
A2. BMP4

Q3. What fold-change do the authors report for the above gene? What fold-change did you calculate for that gene? Please round to 2 decimal places.
A3. 1.22, 1.28

Q4. From the publication text, the individuals that provided samples were all in what stage of the menstrual cycle? Briefly explain why this might be important to consider.
A4. Proliferative

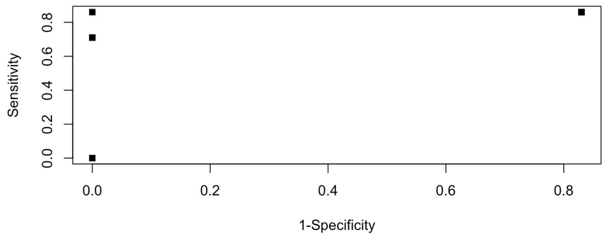
Q5. Briefly explain why the information in question 4 might be important to consider in this analysis.
A5. Expression could change depending on time in the menstrual cycle and could confound results.

Part 2. KNN parameters

These questions will use the code you wrote for KNN.

Q6. Run your K-NN classifier with k=3 over values of fn = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 1. Create an ROC curve from the results and upload here.

A6. (resembles a step function)



Q7. Run your KNN classifier over values of k=1,2,3,4,5,6 with fn = 0.5. List here any samples that are consistently misclassified (>3 times), and the type of misclassification (TP,FP,TN,FN).
A7. GSM629719, false negative

Q8. What is the maximum accuracy achieved from above, with fn = 0.5 and k from 1 to 6? Please report your answer as a decimal, not percent, and round to the nearest hundredth.
A8. 0.92

Q9. Briefly explain under what types of circumstances we would want to optimize for sensitivity over specificity and vice versa.
A9. We are looking for some analysis of relative costs of false positives vs. false negatives. A possible answer could look like:

> It depends on what kind of error you need to avoid: type 1 (false positives) or type 2 (false negatives). For example, if your classifier predicts cancer presence/severity, you'd need good sensitivity if you wanted to detect cancer incidence for follow-up testing (the risk of a false positive means a person goes untreated); if you wanted to send cases to aggressive chemotherapy you'd want high specificity (the risk of a false negative means the treatment was for nothing).

Typically, we want *higher sensitivity* in diagnostic setting, especially if the costs of follow up testing is cheap. For treatment settings, it can be a bit more tricky, where adverse side-effects should be accounted for.

<u>Part 3. These questions will refer to the GSEA analysis.</u>

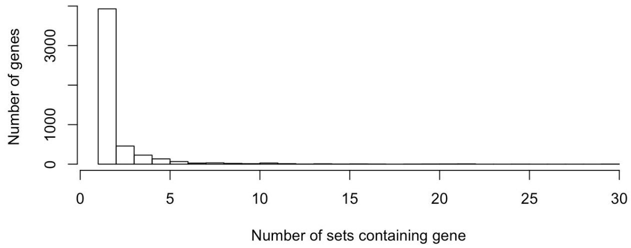
Q10. Which gene set has the highest enrichment score?
A10. KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION

Q11. Please provide a brief biological explanation for the above result
A11. Endometriosis patients are at higher risk of developing MS, evidence of immune modulation through dysregulation of neurotransmitters (observed in literature). Could

also mention a hypothesis of the gene set involved with pain or hormonal pathways (although literature is a bit sparse for these arguments).

Q12. From the kegg file provided, how many unique genes are represented across the provided KEGG sets? Provide just an integer value as your response.
A12. 4991

Q13. Plot a histogram of the number of times each gene appears across the KEGG sets.
A13.



Q13. Which gene(s) appears most often across the KEGG sets? Briefly explain why this might be the case.
A13. MAPK1 and MAPK3

Q14. Provide a brief potential biological explanation as to why the gene(s) listed in the above question might appear the most frequently across gene sets.
A14. MAPKs are kinases involved in many signalling cascades and play roles in many pathways. They are often regarded as "master regulators."

Q15. Briefly discuss limitations of using GSEA to identify genes of interest in a state.
A15. Here are some ideas one can expound upon:

- **GSEA only typically includes well-studied genes and gene sets**
- **Sets can be correlated and some genes appear many times**
- **If one tries enough sets, one will likely be enriched – interpretation isn't clear in these cases**
- **We lose gene-level resolution when binning genes into gene sets; if signal in a given state is highly specific to a particular gene, then GSEA could miss it if that signal isn't strong enough to overcome the lack of signal from other gene set members**