University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Final Exam SOLUTIONS

December 14th, 2023 Name: ____________________

33 Questions, 3 points each (one point is free)

Question 1. Suppose we are doing a pathway enrichment analysis. For a set of 200 DE genes, we calculate one enrichment p-value for each

(A) Gene on the list
(B) Gene Set ⟵ **THIS ONE**
(C) Pair: a gene G on the list and a Gene set
(D) pair of gene sets

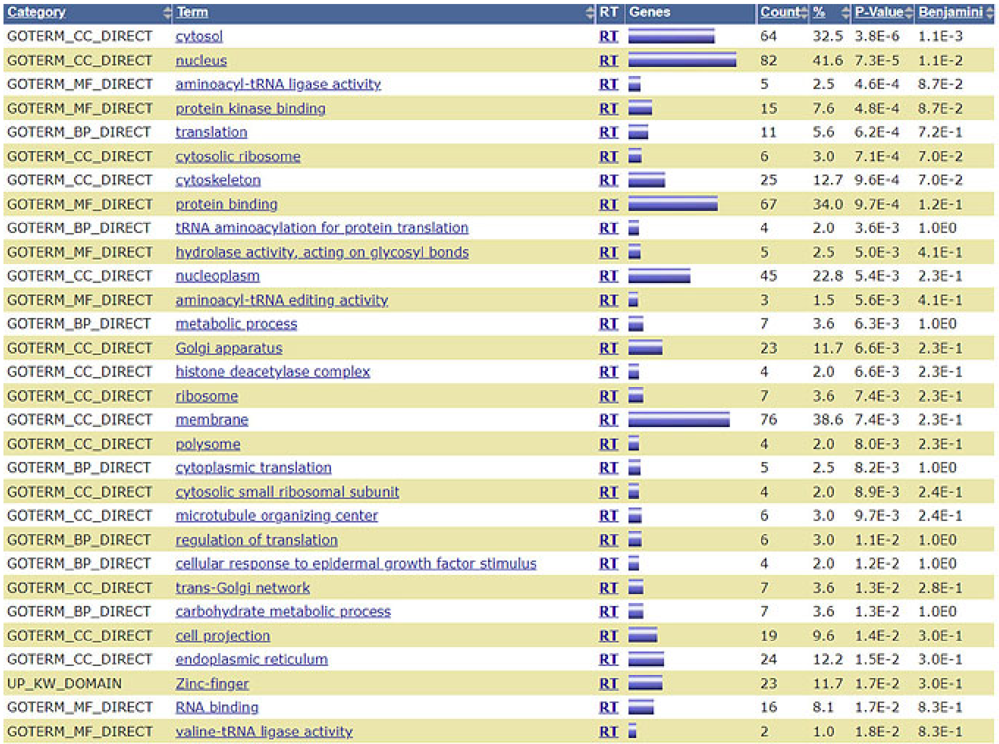
Question 2. True or False. A pathway enrichment analysis p-value is specific to one species.

ANSWER: True.

Question 3. True or False. Input to a pathway analysis is a list of gene identifiers, not a list of isoforms.

ANSWER: True.

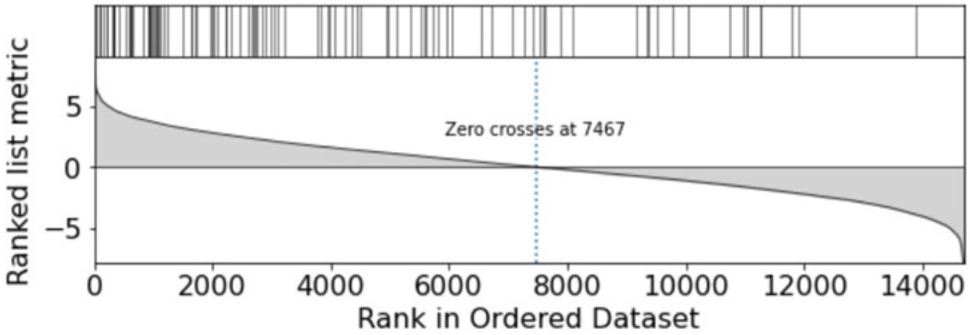
Question 4. Consider the following pathway enrichment results table. Should we consider "cytoplasmic translation" to be significant?



Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_CC_DIRECT	cytosol	RT	▬	64	32.5	3.8E-6	1.1E-3
GOTERM_CC_DIRECT	nucleus	RT	▬	82	41.6	7.3E-5	1.1E-2
GOTERM_MF_DIRECT	aminoacyl-tRNA ligase activity	RT	▬	5	2.5	4.6E-4	8.7E-2
GOTERM_MF_DIRECT	protein kinase binding	RT	▬	15	7.6	4.8E-4	8.7E-2
GOTERM_BP_DIRECT	translation	RT	▬	11	5.6	6.2E-4	7.2E-1
GOTERM_CC_DIRECT	cytosolic ribosome	RT	▬	6	3.0	7.1E-4	7.0E-2
GOTERM_CC_DIRECT	cytoskeleton	RT	▬	25	12.7	9.6E-4	7.0E-2
GOTERM_MF_DIRECT	protein binding	RT	▬	67	34.0	9.7E-4	1.2E-1
GOTERM_BP_DIRECT	tRNA aminoacylation for protein translation	RT	▬	4	2.0	3.6E-3	1.0E0
GOTERM_MF_DIRECT	hydrolase activity, acting on glycosyl bonds	RT	▬	5	2.5	5.0E-3	4.1E-1
GOTERM_CC_DIRECT	nucleoplasm	RT	▬	45	22.8	5.4E-3	2.3E-1
GOTERM_MF_DIRECT	aminoacyl-tRNA editing activity	RT	▬	3	1.5	5.6E-3	4.1E-1
GOTERM_BP_DIRECT	metabolic process	RT	▬	7	3.6	6.3E-3	1.0E0
GOTERM_CC_DIRECT	Golgi apparatus	RT	▬	23	11.7	6.6E-3	2.3E-1
GOTERM_CC_DIRECT	histone deacetylase complex	RT	▬	4	2.0	6.6E-3	2.3E-1
GOTERM_CC_DIRECT	ribosome	RT	▬	7	3.6	7.4E-3	2.3E-1
GOTERM_CC_DIRECT	membrane	RT	▬	76	38.6	7.4E-3	2.3E-1
GOTERM_CC_DIRECT	polysome	RT	▬	4	2.0	8.0E-3	2.3E-1
GOTERM_BP_DIRECT	cytoplasmic translation	RT	▬	5	2.5	8.2E-3	1.0E0
GOTERM_CC_DIRECT	cytosolic small ribosomal subunit	RT	▬	4	2.0	8.9E-3	2.4E-1
GOTERM_CC_DIRECT	microtubule organizing center	RT	▬	6	3.0	9.7E-3	2.4E-1
GOTERM_BP_DIRECT	regulation of translation	RT	▬	6	3.0	1.1E-2	1.0E0
GOTERM_BP_DIRECT	cellular response to epidermal growth factor stimulus	RT	▬	4	2.0	1.2E-2	1.0E0
GOTERM_CC_DIRECT	trans-Golgi network	RT	▬	7	3.6	1.3E-2	2.8E-1
GOTERM_BP_DIRECT	carbohydrate metabolic process	RT	▬	7	3.6	1.3E-2	1.0E0
GOTERM_CC_DIRECT	cell projection	RT	▬	19	9.6	1.4E-2	3.0E-1
GOTERM_CC_DIRECT	endoplasmic reticulum	RT	▬	24	12.2	1.5E-2	3.0E-1
UP_KW_DOMAIN	Zinc-finger	RT	▬	23	11.7	1.7E-2	3.0E-1
GOTERM_MF_DIRECT	RNA binding	RT	▬	16	8.1	1.7E-2	8.3E-1
GOTERM_MF_DIRECT	valine-tRNA ligase activity	RT	▬	2	1.0	1.8E-2	8.3E-1

ANSWER: No, because the q-value equals one.

Question 5. In the following GSEA diagram, the black vertical lines at the top represent
(A) The genes in the gene set of interest ⟵ **THIS ONE**
(B) The genes outside the gene set of interest
(C) The DE genes
(D) The SNP locations that are eQTL's
(E) Indels



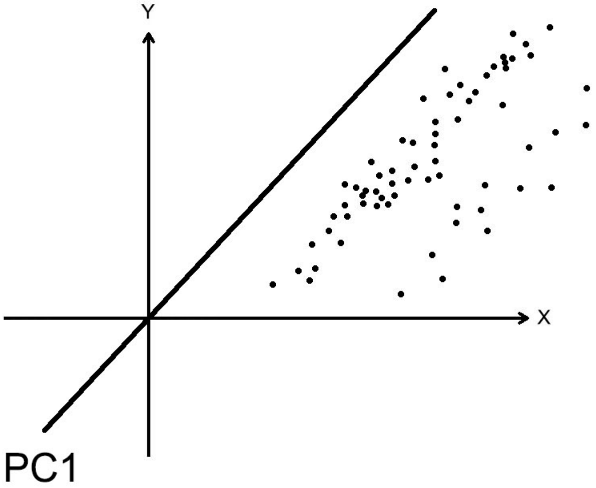
Question 6. True or False. A "subspace" of n-dimensional space must contain the origin.

ANSWER: True. By definition, a subspace must contain the origin.

Question 7. True or False. In Principle Components Analysis, the first principle component PC1 captures the technical variation and the second principle component PC2 captures the biological variation.

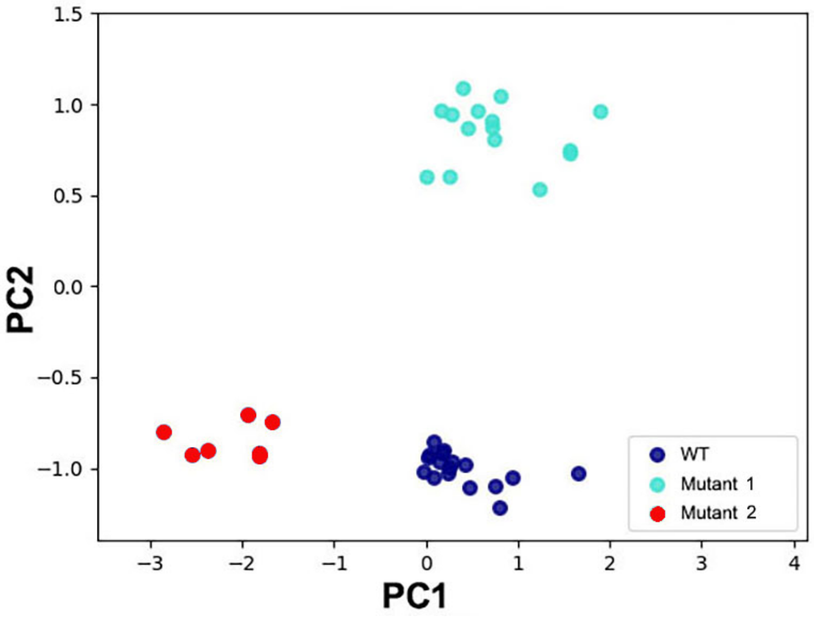
ANSWER: False, which is PC1 it depends on which factor is responsible more variance.

Question 8. On the following graph, draw in (approximately) the line correponding to the first principle component subspace PC1.



ANSWER: It's the direction of greatest variance, not the direction of greatest separation between the two apparent clusters.

Question 9. Suppose you have RNA-Seq data from three experimental conditions WT, Mutant 1 and Mutant 2 and you get the following PCA plot. Suppose the loadings for PC1 are non-zero only in pathway P_1 and the loadings for PC2 are non-zero only in pathway P_2. Interpret the following PCA plot.



ANSWER: Pathway P_1 is differentially expressed between WT and Mutant 2 and pathway P_2 is differentially expressed between WT and Mutant 1. In other words the difference between WT and Mutant 2 is explained by pathway P_1 and the difference between WT and Mutant 1 is explained by pathway P_2

Question 10. True or False. A Mann-Whitney test cannot declare significane of a comparison where there are two replicates per group, no matter what the data.

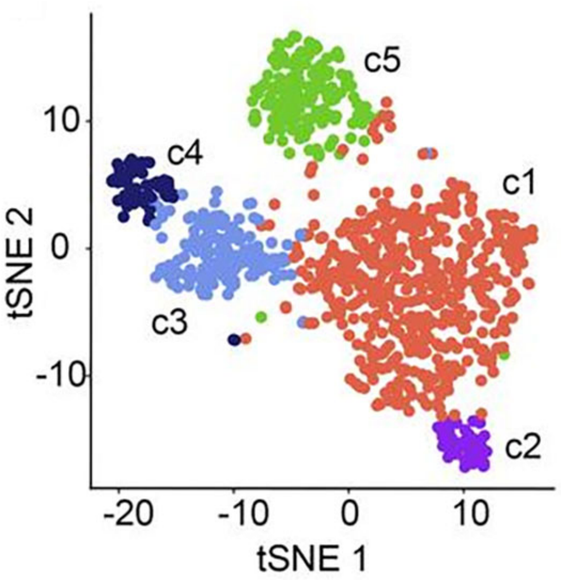
ANSWER: True.

Question 11. The Mann-Whitney test is robust to outliers because (circle the one correct answer)
(A) It uses a normal distribution which has thin tails.
(B) Because it requires a lot of replicates, so outliers are negligible.
(C) It is based on ranking, which is blind to outliers. ⟵ **THIS ONE**
(D) Because Mann-Whitney is a permutation test.

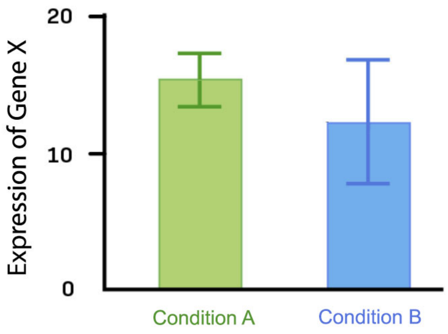
Question 12. True or False. Suppose a permutation p-value is calculated using all permutations. Let N is the total number of permutations. The smallest the permutation p-value can be is $1/N$.

ANSWER: True.

Question 13. In the following single cell RNA-Seq tSNE plot, each point represents: (circle one)
(A) One gene
(B) One subject
(C) One pathway
(D) One significance level
(E) One cell ⟵ **THIS ONE**

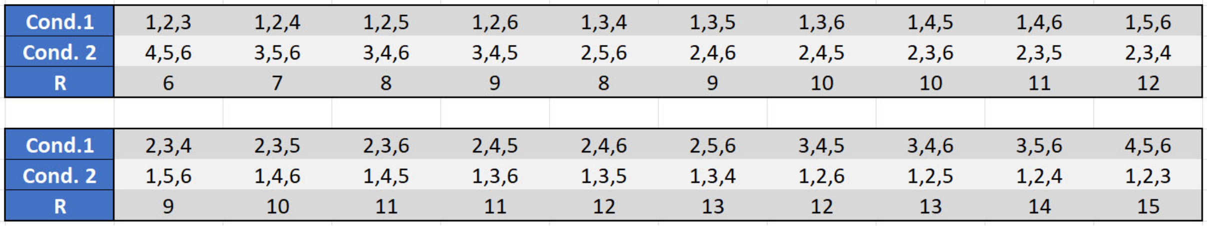


Question 14. Consider the data in the following graph of expression of Gene X between Condition A and Condition B. Explain why we should not apply a parametric T-test here.



ANSWER: Because there's much greater variance in Condition B and a parametric T-test requires them to be equal.

Question 15. The table below shows all 20 possible rankings of a 3-versus-3 comparision for a Mann-Whitney analysis. The table is split over two rows since it was too wide to display on one. Each ranking is equally likely, so each has probability $1/20 = 0.05$. What is the probability that $R = 9$? In other words, what is $P(R = 9)$?



Cond.1	1,2,3	1,2,4	1,2,5	1,2,6	1,3,4	1,3,5	1,3,6	1,4,5	1,4,6	1,5,6
Cond. 2	4,5,6	3,5,6	3,4,6	3,4,5	2,5,6	2,4,6	2,4,5	2,3,6	2,3,5	2,3,4
R	6	7	8	9	8	9	10	10	11	12

Cond.1	2,3,4	2,3,5	2,3,6	2,4,5	2,4,6	2,5,6	3,4,5	3,4,6	3,5,6	4,5,6
Cond. 2	1,5,6	1,4,6	1,4,5	1,3,6	1,3,5	1,3,4	1,2,6	1,2,5	1,2,4	1,2,3
R	9	10	11	11	12	13	12	13	14	15

ANSWER: $R = 9$ for three different rankings, each has probability $1/20$ so the probability $R = 9$ is $3/20$.

Question 16. For the majority of GWAS studies, SNP calling is done with (choose one)
(A) Microarrays ⟵ **THIS ONE**
(B) DNA-Seq
(C) PCR

Question 17. Every point on a Manhattan plot represents (choose one)
(A) One subject
(B) One phenotype
(C) One SNP ⟵ **THIS ONE**
(D) One codon

Question 18. True or False. Fine Mapping refers to finding the exact location of the most significant SNP in a given locus.

ANSWER: False, it's not about the most significant SNP, it's about finding the causative SNP.

Question 19. In a manhattan plot explain the rationale behind why we graph the Y-axis as $-\log_{10}(p)$ and not just p (where p is the p-value).

ANSWER: Because it makes the p-value cutoffs 0.1, 0.01, 0.001, etc. equally spaced along the axis.

Question 20. A "polygenic risk score" is (circle all that apply):
(A) Used for assesing disease risk ⟵ **THIS ONE**
(B) Is based on multiple SNPs ⟵ **THIS ONE**
(C) Is used to infer mechanism of action
(D) Might be found in a person's medical chart ⟵ **THIS ONE**
(E) Is based on gene expression.

Question 21. True or False. Supervised learning is about prediction and unsuperviesd learning is about classification.

ANSWER: False. Classification and prediction are performed by supervised learning.

Question 22. In the Learning Inequality, why is it uniformative when the hypothesis set $\mathcal{H}$ consists of all straight lines in the plane?

$$\boxed{P\left(\left|E_{in}(h)-E_{out}(h)\right|>\varepsilon\right) \leqslant 2|\mathscr{H}|\cdot e^{-2n\varepsilon^2}}$$

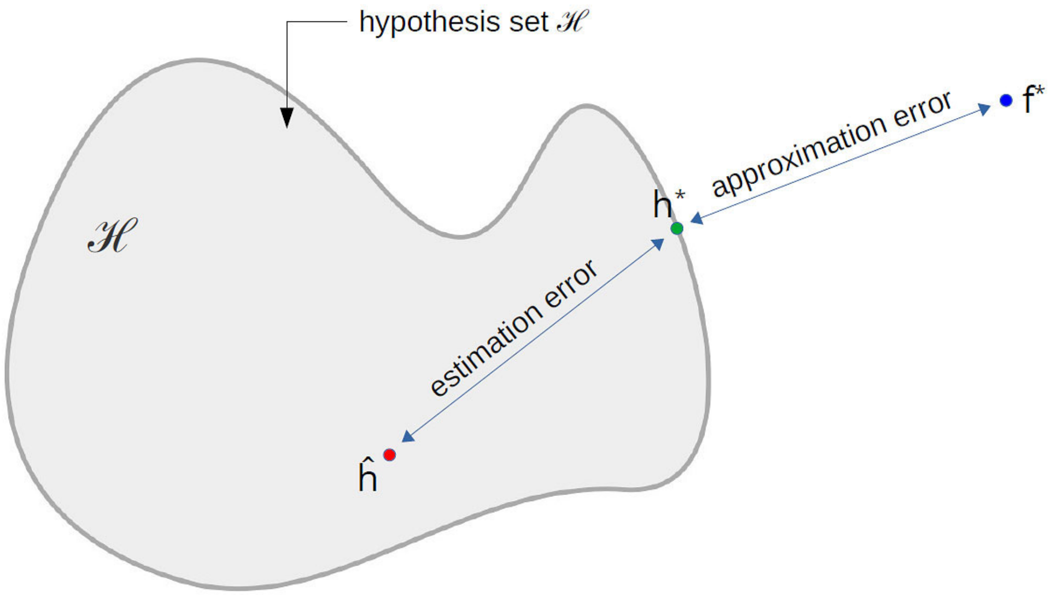
ANSWER: Because then $\mathcal{H}$ is infinite, so $|\mathcal{H}| = \infty$ and everything is less or equal to infinity.

Question 23. In supervised learning regression, what is learned from the trianing data?
(A) The form of the model
(B) The parameters of the model ⟵ **THIS ONE**
(C) The hypothesis set
(D) The test data

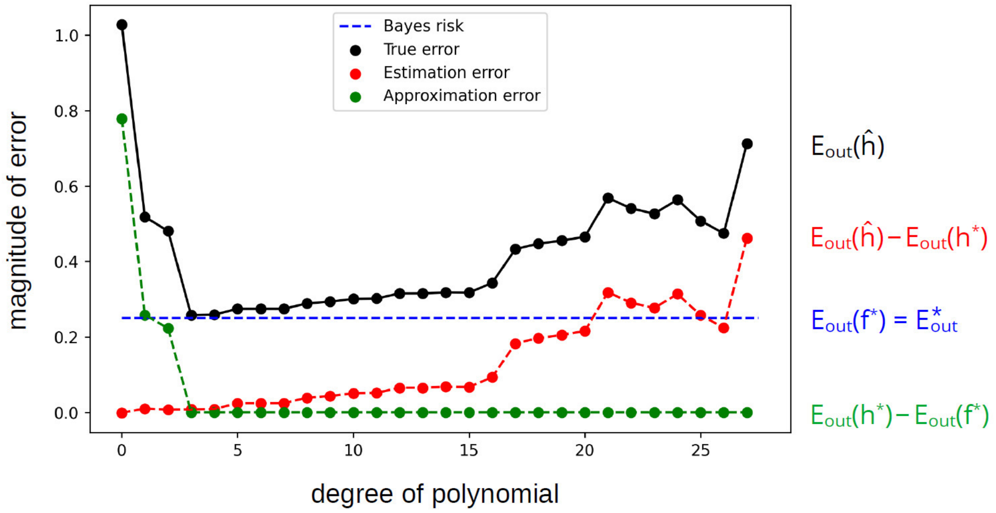
Question 24. In the figure, which error is due to overfitting?

(A) Estimation Error ⟵ **THIS ONE**

(B) Approximation Error

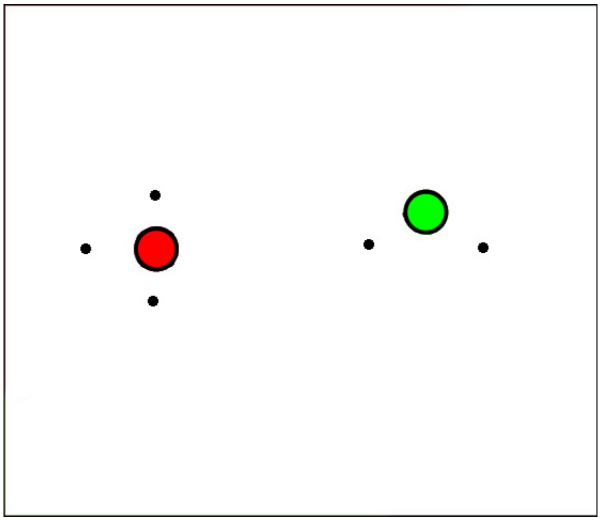


Question 25. True or False. The green points decrease to almost zero but can never be exactly zero.

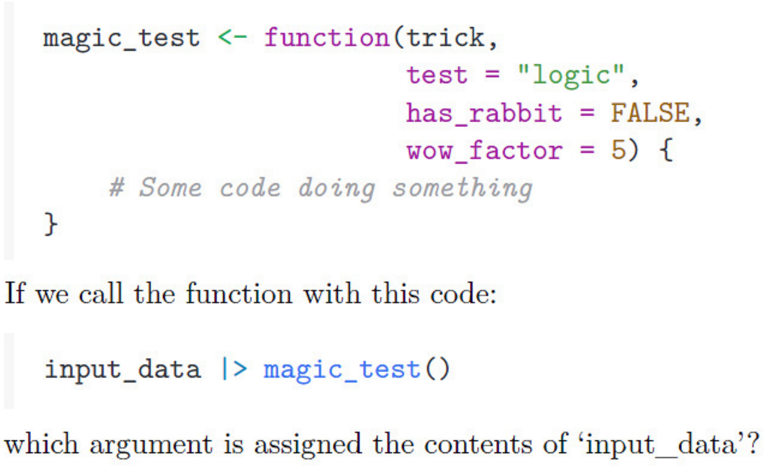


ANSWER: False, they decrease to zero once the true model is contained in $\mathcal{H}$.

Question 26. The figure shows k-means clustering with five data points and $k = 2$. On the next iteration of the algorithm, one of these is correct, which one?
(A) Cluster membership and the centroid locations will change.
(B) Cluster membership will change but the centroid locations will not change.
(C) Cluster membership will not change, but the centroid locations will change. ⟵ **THIS ONE**
(D) Neither cluster membership, nor the centroid locations will change.



Question 27. Consider the following definition for the `magic_test` function:



```r
magic_test <- function(trick,
                       test = "logic",
                       has_rabbit = FALSE,
                       wow_factor = 5) {
    # Some code doing something
}
```

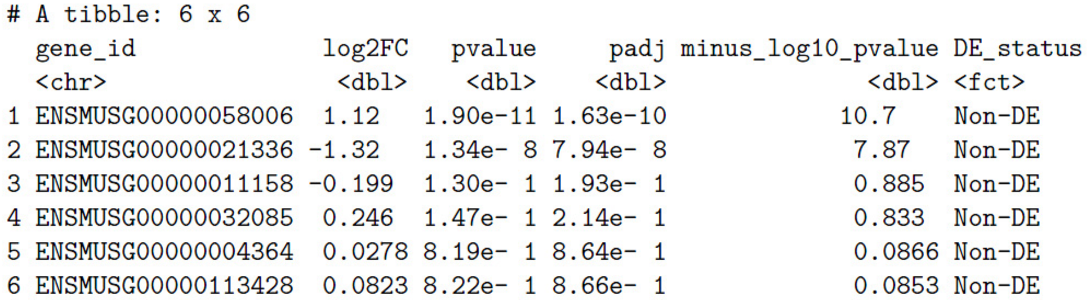
If we call the function with this code:

```r
input_data |> magic_test()
```

which argument is assigned the contents of 'input_data'?

(A) `trick` ⟵ **THIS ONE**
(B) `test`
(C) `has_rabbit`
(D) `wow_factor`

Question 28. Consider the 'de_results' data frame of differential expression results:



```
# A tibble: 6 x 6
  gene_id             log2FC   pvalue     padj minus_log10_pvalue DE_status
  <chr>                <dbl>    <dbl>    <dbl>              <dbl> <fct>
1 ENSMUSG00000058006  1.12   1.90e-11 1.63e-10            10.7    Non-DE
2 ENSMUSG00000021336 -1.32   1.34e- 8 7.94e- 8             7.87   Non-DE
3 ENSMUSG00000011158 -0.199  1.30e- 1 1.93e- 1             0.885  Non-DE
4 ENSMUSG00000032085  0.246  1.47e- 1 2.14e- 1             0.833  Non-DE
5 ENSMUSG00000004364  0.0278 8.19e- 1 8.64e- 1             0.0866 Non-DE
6 ENSMUSG00000113428  0.0823 8.22e- 1 8.66e- 1             0.0853 Non-DE
```

Which R code would sort the 'de_results' data frame by the values in the log2FC column, from largest to smallest?

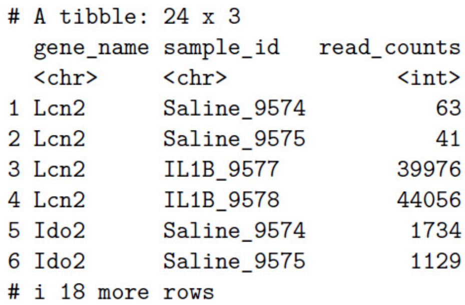
(A) `arrange(log2FC, de_results)`
(B) `arrange(desc(log2FC), de_results)`
(C) `arrange(de_results, desc(log2FC))` ⟵ **THIS ONE**
(D) `arrange(de_results, log2FC)`

Question 29. Fill in the blank with the correct R operator to assign the value of 0.01 to the variable `deg_cutoff`

`deg_cutoff <- 0.01`

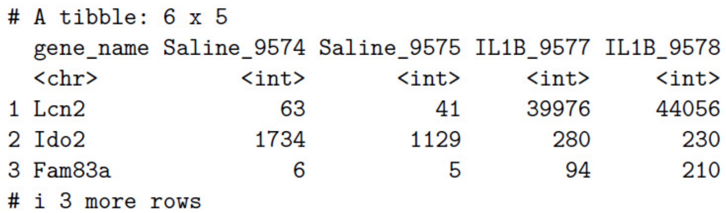
Question 30. Consider these two tibbles:

Tibble A:



```
# A tibble: 24 x 3
  gene_name sample_id   read_counts
  <chr>     <chr>             <int>
1 Lcn2      Saline_9574          63
2 Lcn2      Saline_9575          41
3 Lcn2      IL1B_9577         39976
4 Lcn2      IL1B_9578         44056
5 Ido2      Saline_9574        1734
6 Ido2      Saline_9575        1129
# i 18 more rows
```

Tibble B:

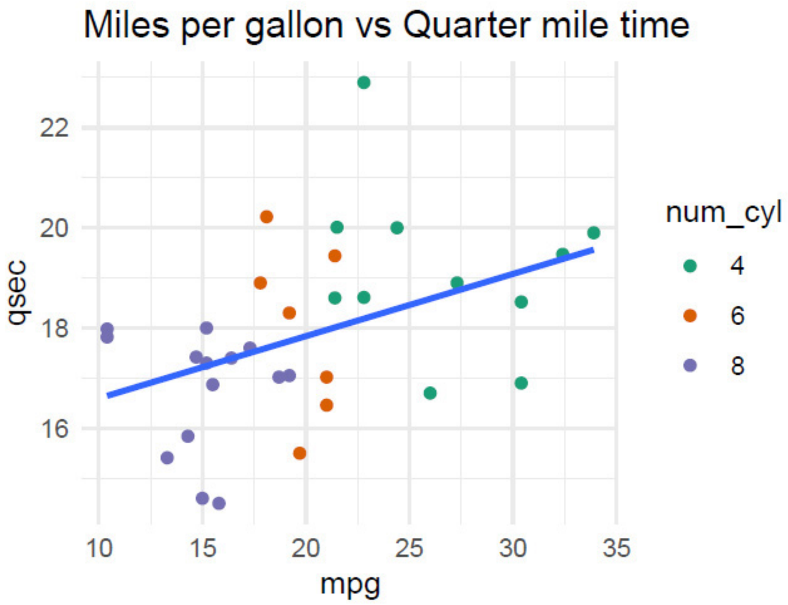


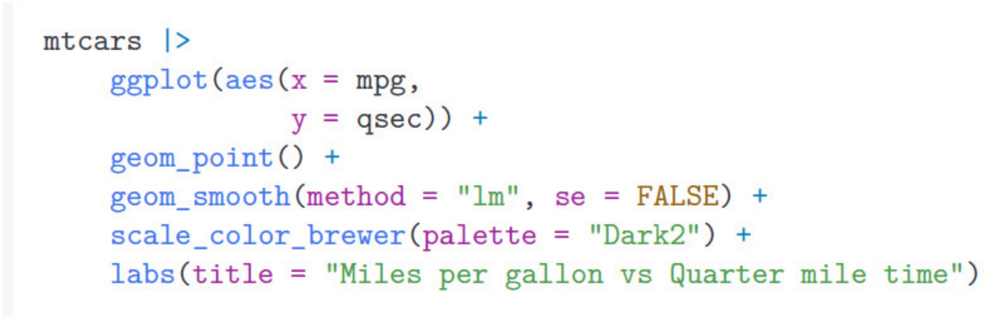
```
# A tibble: 6 x 5
  gene_name Saline_9574 Saline_9575 IL1B_9577 IL1B_9578
  <chr>           <int>       <int>     <int>     <int>
1 Lcn2               63          41     39976     44056
2 Ido2             1734        1129       280       230
3 Fam83a              6           5        94       210
# i 3 more rows
```

Which R function would you use to reshape Tibble B into Tibble A?

(A) `left_join()`
(B) `pivot_longer()` ⟵ **THIS ONE**
(C) `pivot_wider()`
(D) `bind_rows()`

Question 31. Consider the following graph:





```
mtcars |>
    ggplot(aes(x = mpg,
               y = qsec)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    scale_color_brewer(palette = "Dark2") +
    labs(title = "Miles per gallon vs Quarter mile time")
```

In which *ggplot2* function would you add the color '`color=num_cyl`' aesthetic mapping to recreate this graph?

(A) ggplot()
(B) geom_point() ⟵ **THIS ONE**
(C) geom_smooth()
(D) scale_color_brewer()

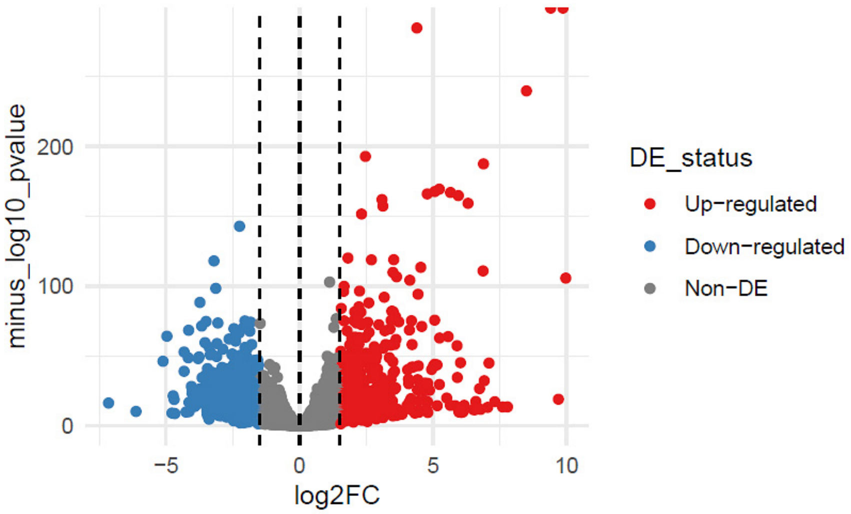
Question 32. Circle the R pipe operator

(A) <=
(B) | > ⟵ **THIS ONE**
(C) < −
(D) ==

Question 33. Consider the ’de_results’ data frame of differential expression results:

```
# A tibble: 6 x 6
  gene_id             log2FC   pvalue     padj minus_log10_pvalue DE_status
  <chr>                <dbl>    <dbl>    <dbl>              <dbl> <fct>
1 ENSMUSG00000058006  1.12   1.90e-11 1.63e-10            10.7    Non-DE
2 ENSMUSG00000021336 -1.32   1.34e- 8 7.94e- 8             7.87   Non-DE
3 ENSMUSG00000011158 -0.199  1.30e- 1 1.93e- 1             0.885  Non-DE
4 ENSMUSG00000032085  0.246  1.47e- 1 2.14e- 1             0.833  Non-DE
5 ENSMUSG00000004364  0.0278 8.19e- 1 8.64e- 1             0.0866 Non-DE
6 ENSMUSG00000113428  0.0823 8.22e- 1 8.66e- 1             0.0853 Non-DE
```

Here’s a volcano plot made from the ’de_results’ data frame:



Which `geom_` function(s) would you need to create this volcano plot (circle all that apply).

(A) `geom_hline()`
(B) `geom_vline()` ⟵ **THIS ONE**
(C) `geom_violin()`
(D) `geom_point()` ⟵ **THIS ONE**