This is not a practice exam. It is a rewrite of some of the lab questions in an exam format.

1. Adapted from Lab 1A Smoking

A study of patients with insulin-dependent diabetes was conducted to investigate the effects of cigarette smoking on renal and retinal complications. Before examining the results of the study, a researcher expects that the proportions of four different subgroups are as follow:

Subgroup	Proportion
Nonsmokers	0.50
Current Smokers	0.20
Tobacco Chewers	0.10
Ex-smokers	0.20

Of 100 randomly selected patients, there are 44 nonsmokers, 24 current smokers, 13 tobacco chewers and 19 ex-smokers. Should the researcher revise his estimates? Use 0.01 level of significance.

(a) Which test is most appropriate in this scenario?

(b) Write down the appropriate null and alternative hypotheses.

(c) What are the assumptions required for this test. Are they satisfied here?

(d) What is the approximate distribution of the test statistic under the null hypothesis?

(e) Write down an expression for the p-value.

(f) What is your decision for the test and why?

```
> y_i = c(44, 24, 13, 19)
> p_i = c(0.5, 0.2, 0.1, 0.2)
> (n = sum(y_i))

[1] 100

> (e_i = n * p_i)

[1] 50 20 10 20

> sum((y_i - e_i)^2/e_i)

[1] 2.47

> qchisq(0.005, 1:6, lower.tail = FALSE)

[1]  7.879439 10.596635 12.838156 14.860259 16.749602 18.547584

> qchisq(0.01, 1:6, lower.tail = FALSE)

[1]  6.634897  9.210340 11.344867 13.276704 15.086272 16.811894

> qchisq(0.025, 1:6, lower.tail = FALSE)

[1]  5.023886  7.377759  9.348404 11.143287 12.832502 14.449375

> qchisq(0.05, 1:6, lower.tail = FALSE)

[1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587
```

2. **Adapted from Lab 1B Mammograms**

Suppose that among 100,000 women with negative mammograms, 20 will have breast cancer diagnosed within 2 years; and among 100 women with positive mammograms, 10 will have breast cancer diagnosed within 2 years. Clinicians would like to know if there is a relationship between a positive or negative mammogram and developing breast cancer?

Mammogram Breast cancer	Yes	No
Positive	10	90
Negative	20	99,980

(a) Is this a retrospective or a prospective study?

(b) Is it appropriate to use a relative risk to quantify the relationship between the risk factor (mammogram result) and disease (breast cancer)? If so calculate the relative risk.

(c) Calculate the odds of having breast cancer for positive vs negative mammograms.

(d) Calculate the standard error for the log odds-ratio.

(e) Calculate a 95% confidence interval for the odds-ratio.

(f) Is there evidence that there might be a relationship between mammogram test results and breast cancer diagnosis?

Some R output that may be helpful:

```
> qnorm(c(0.9,0.95,0.975))

[1] 1.281552 1.644854 1.959964

> qchisq(c(0.9, 0.95, 0.975), 1)

[1] 2.705543 3.841459 5.023886

> qt(c(0.9, 0.95, 0.975), 4)

[1] 1.533206 2.131847 2.776445
```

3. **Adapted from Lab 1C TV violence**

A study of the amount of violence viewed on television as it relates to the age of the viewer yields the results shown in the accompanying table for 81 people.

	Age		
Viewing	$16-34$	$35-54$	55 and over
Low violence	8	12	21
High violence	18	15	7

(a) Which test is most appropriate in this scenario?

(b) Write down the appropriate null and alternative hypotheses.

(c) What are the assumptions required for this test. Are they satisfied here?

(d) What is the approximate distribution of the test statistic under the null hypothesis?

(e) Write down an expression for the p-value.

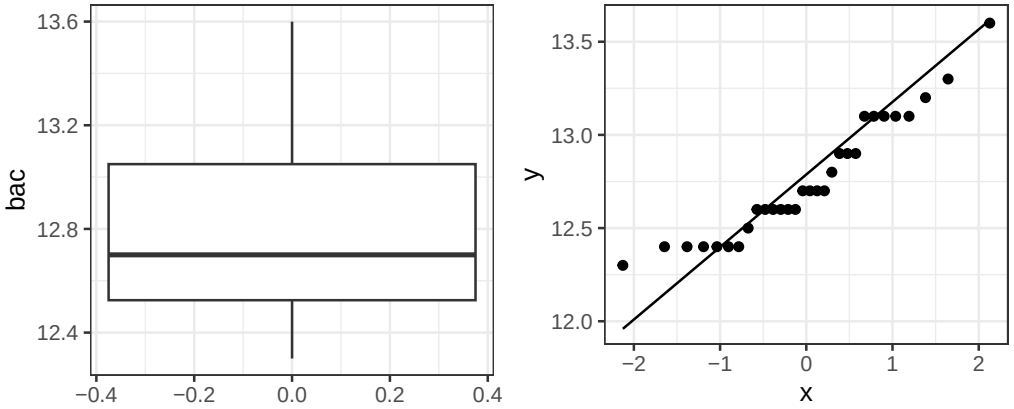
(f) What is your decision for the test and why?

```
> x = matrix(c(8, 18, 12, 15, 21, 7), ncol = 3)
> colnames(x) = c("16-34", "35-54", "54+")
> rownames(x) = c("Low violence", "High violence")
> x

              16-34 35-54 54+
Low violence      8    12  21
High violence    18    15   7

> (n = sum(x))

[1] 81

> (xr = apply(x, 1, sum))

 Low violence High violence
           41            40

> (xc = apply(x, 2, sum))

16-34 35-54   54+
   26    27    28

> (ex = xr %*% t(xc) / n)

        16-34    35-54      54+
[1,] 13.16049 13.66667 14.17284
[2,] 12.83951 13.33333 13.82716

> sum((x - ex)^2 / ex)

[1] 11.16884

> qchisq(0.05, 1:6, lower.tail = FALSE)

[1]  3.841459  5.991465  7.814728  9.487729 11.070498 12.591587

> qt(0.05, 1:6)

[1] -6.313752 -2.919986 -2.353363 -2.131847 -2.015048 -1.943180

> qt(0.025, 1:6)

[1] -12.706205  -4.302653  -3.182446  -2.776445  -2.570582  -2.446912
```

4. Adapted from Lab 2A Blood alcohol readings

The following are 30 blood alcohol determinations made by Analyzer GTE-10, a three-year-old unit that may be in need of re-calibration. All 30 measurements were made using a test sample on which a properly adjusted machine would give a reading of 12.6%.

```
> bac = c(12.3,12.7,12.6,13.1,13.2,12.8,13.1,12.9,13.1,12.4,13.6,12.7,12.6,13.1,12.4,
+         12.6,13.3,12.6,12.4,13.1,12.9,12.6,12.7,12.5,12.4,12.4,12.6,12.7,12.4,12.9)

> library(tidyverse)
> bac_df = data.frame(bac)
> p1 = ggplot(bac_df, aes(y = bac)) + geom_boxplot() + theme_bw()
> p2 = ggplot(bac_df, aes(sample = bac)) + geom_qq() + geom_qq_line() + theme_bw()
> gridExtra::grid.arrange(p1,p2, ncol = 2)
```




```
> n = length(bac)
> xbar = mean(bac)
> s = sd(bac)
> c(n, xbar, s)

[1] 30.0000000 12.7566667  0.3244978

> qnorm(c(0.9,0.95,0.975),29)

[1] 30.28155 30.64485 30.95996

> qt(c(0.9,0.95,0.975),29)

[1] 1.311434 1.699127 2.045230

> qt(c(0.9,0.95,0.975),30)

[1] 1.310415 1.697261 2.042272
```

(a) Write out the hypothesis for the Analyzer GTE-10 being faulty.

(b) What are the assumptions of this test? Are they satisfied?

(c) Assuming that the assumptions are satisfied (regardless of what you found above), write down the test statistic and its distribution assuming that the null hypothesis is true.

(d) Calculate the observed test statistic.

(e) At the level of significance $\alpha = 0.05$, what is your conclusion?

5. **Adapted from Lab 2B Weight gain**

10 pigs were independently sampled and fed a specific diet. The weight of 5 pigs on diet X and 5 pigs on diet Y are

Diet X: 12, 16, 16, 12, 10 and diet Y: 30, 12, 24, 32, 24.

We want to test if there is a difference in weight between the two diets using the Wilcoxon rank-sum test.

(a) Write out the null and alternative hypotheses. [Be sure to define all parameters used.]

(b) Calculate the Wilcoxon rank-sum test statistic and the standardised version of the test statistic.

(c) At the level of significance $\alpha = 0.05$, what is your conclusion?

(d) What is a parametric test that could be used instead of the Wilcoxon rank-sum test? What is the advantage of using a Wilcoxon rank-sum test over a parametric test? What would be the advantage of using a parametric test over the Wilcoxon rank-sum test?

(e) Describe how a permutation test works. How would you implement a permutation test in this context?

```
> wdat = data.frame(
+   diet = rep(c("X","Y"), each = 5),
+   weight = c(12, 16, 16, 12, 10, 30, 12, 24, 32, 24)
+   ) %>%
+   mutate(ranks = rank(weight))
> wdat

   diet weight ranks
1     X     12   3.0
2     X     16   5.5
3     X     16   5.5
4     X     12   3.0
5     X     10   1.0
6     Y     30   9.0
7     Y     12   3.0
8     Y     24   7.5
9     Y     32  10.0
10    Y     24   7.5

> wdat %>% group_by(diet) %>% summarise(sum(ranks))

# A tibble: 2 × 2
  diet  `sum(ranks)`
  <chr>        <dbl>
1 X               18
2 Y               37

> nx = 5
> ny = 5
> N = nx + ny
> ew = nx*(N+1)/2
> varw = (sum(wdat$ranks^2) - N*(N+1)^2/4)*nx*ny/(N*(N-1))
> c(ew, varw)

[1] 27.50000 22.08333

> qnorm(c(0.9,0.95,0.975))

[1] 1.281552 1.644854 1.959964

> qt(c(0.9,0.95,0.975), 8)

[1] 1.396815 1.859548 2.306004
```

6. **Adapted from Lab 3A Flicker frequency**

If a light is flickering but at a very high frequency, it appears to not be flickering at all. Thus there exists a "critical flicker frequency" where the flickering changes from "detectable" to "not detectable" and this varies from person to person.

The critical flicker frequency and iris colour for 19 randomly sampled people were obtained as part of a study into the relationship between critical frequency flicker and eye colour.

We want to use a one-way ANOVA to test if there is a significant difference in the mean detectable flicker frequency between people with different eye colours.

(a) Write out the appropriate null and alternative hypotheses. [Be sure to define all parameters used.]

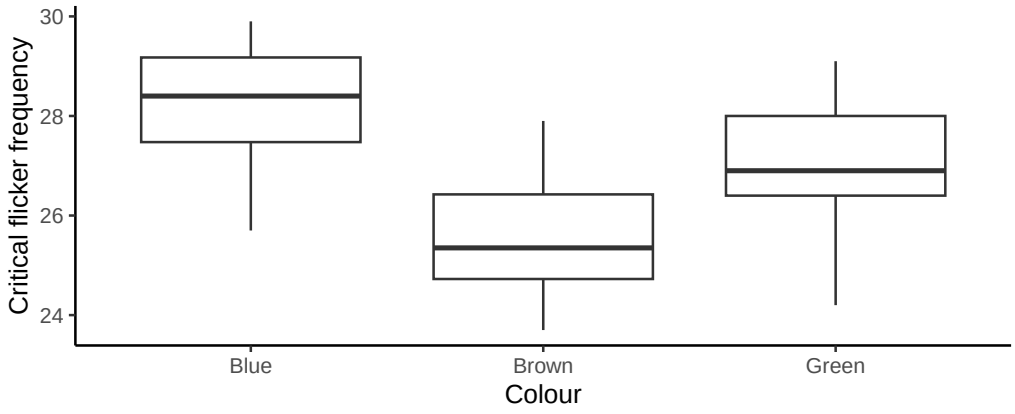
(b) What are the assumptions required for a one-way ANOVA? Are they satisfied in this case?

(c) Using the output, write out the hypothesis test in full. Be sure to state the null and alternative hypothesis, assumptions, test statistic (with distribution), observed test statistic, p-value and an appropriate conclusion.

(d) If appropriate, discuss the post hoc test results to identify which pairwise differences are significant. If not appropriate, give a brief justification as to why not.

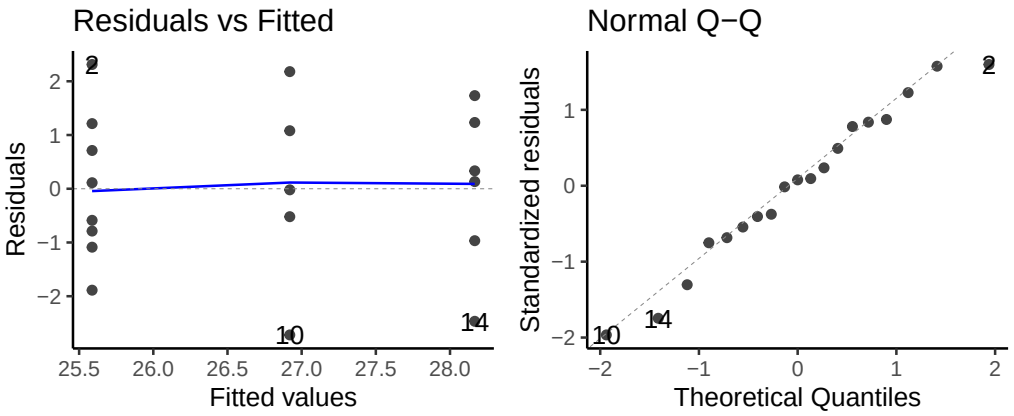
(e) Describe how to perform the Bonferroni correction in the context of post-hoc pairwise testing. Why is it needed?

```
> library(tidyverse)
> flicker = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/flicker.txt")
> glimpse(flicker)

Rows: 19
Columns: 2
$ Colour  <chr> "Brown", "Brown", "Brown", "Brown", "Brown", "Brown", "Br...
$ Flicker <dbl> 26.8, 27.9, 23.7, 25.0, 26.3, 24.8, 25.7, 24.5, 26.4, 24....

> ggplot(flicker, aes(x = Colour, y = Flicker)) + geom_boxplot() + theme_classic() +
+   labs(y = "Critical flicker frequency", y = "Eye colour")
```




```
> flicker_anova = aov(Flicker ~ Colour, data = flicker)
> summary(flicker_anova)

            Df Sum Sq Mean Sq F value Pr(>F)
Colour       2  23.00  11.499   4.802 0.0232
Residuals   16  38.31   2.394

> library(emmeans)
> flicker_emmeans = emmeans(flicker_anova, ~ Colour)
> contrast(flicker_emmeans, method = "pairwise", adjust = "bonferroni")

 contrast      estimate    SE df t.ratio p.value
 Blue - Brown      2.58 0.836 16   3.086  0.0212
 Blue - Green      1.25 0.937 16   1.331  0.6060
 Brown - Green    -1.33 0.882 16  -1.511  0.4512

P value adjustment: bonferroni method for 3 tests
```

```
> library(ggfortify)
> autoplot(flicker_anova, which = c(1,2)) + theme_classic()
```



7. Adapted from Lab 4A Wind

The data in 'pollut.txt' are WS (wind speeds), Temp (temperature), H (humidity), In (insolation) and O (ozone) for 30 days. R output is given below to help you answer the following questions.

(a) Does it look like any variables can be dropped from the model? If you were doing backwards selection using the 'drop1()' function which would you drop first?

(b) Write down a the workflow for a formal hypothesis test to see if the coefficient for insolation is significantly different to zero. Make sure you state the null and alternative hypotheses, test statistic (and its distribution), p-value and conclusion.

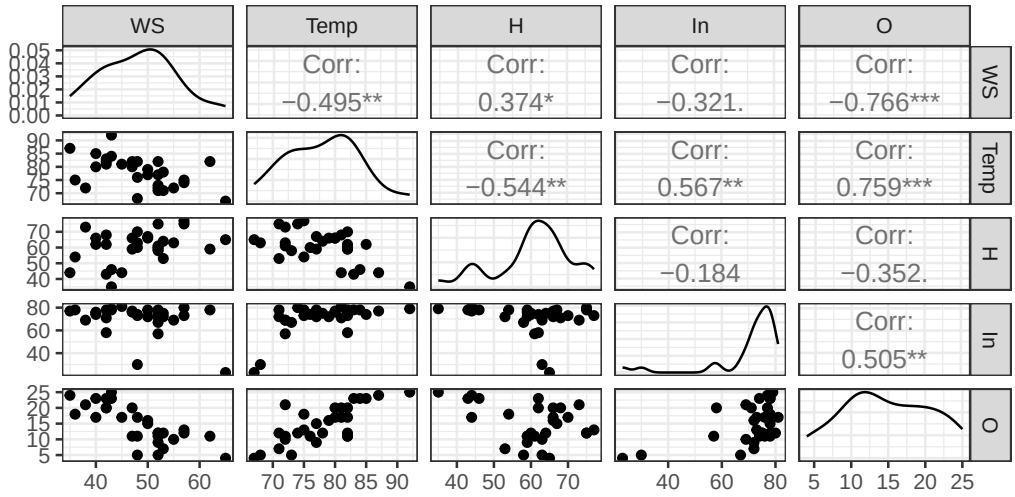
(c) Write down the fitted model for the model selected by the backward stepwise procedure.

(d) State and check the linear regression assumptions for the model selected by the backward stepwise procedure.

(e) What proportion of the variability of ozone is explained by the explanatory variables in the stepwise selected model?

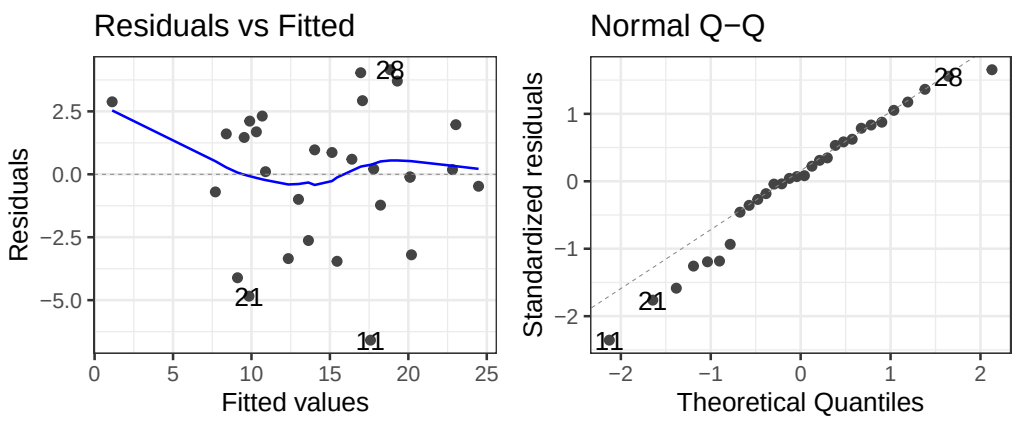
(f) Use the model to estimate the average 'ozone' for days when 'WS=40', 'Temp=80' and 'H=50'. Is a confidence interval or a prediction interval most appropriate here? Write down the interval you think is most appropriate.

```
> pollut = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/pollut.txt")
> glimpse(pollut)

Rows: 30
Columns: 5
$ WS   <dbl> 50, 47, 57, 38, 52, 57, 53, 62, 52, 42, 47, 40, 42, 40, 48, ...
$ Temp <dbl> 77, 80, 75, 72, 71, 74, 78, 82, 82, 82, 82, 80, 81, 85, 82, ...
$ H    <dbl> 67, 66, 77, 73, 75, 75, 64, 59, 60, 62, 59, 66, 68, 62, 70, ...
$ In   <dbl> 78, 77, 73, 69, 78, 80, 75, 78, 75, 58, 76, 76, 71, 74, 73, ...
$ O    <dbl> 15, 20, 13, 21, 12, 12, 12, 11, 12, 20, 11, 17, 20, 23, 17, ...

> library(GGally)
> ggpairs(pollut) + theme_bw()
```




```
> pollut_lm = lm(O ~ ., pollut)
> summary(pollut_lm)

Call:
lm(formula = O ~ ., data = pollut)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5861 -1.0961  0.3512  1.7570  4.0712

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.49370   13.50647  -1.147  0.26219
WS           -0.44291    0.08678  -5.104 2.85e-05
Temp          0.56933    0.13977   4.073  0.00041
H             0.09292    0.06535   1.422  0.16743
In            0.02275    0.05067   0.449  0.65728

Residual standard error: 2.92 on 25 degrees of freedom
Multiple R-squared:  0.798,        Adjusted R-squared:  0.7657
F-statistic: 24.69 on 4 and 25 DF,  p-value: 2.279e-08

> pollut_step = step(pollut_lm, trace = FALSE)
> summary(pollut_step)

Call:
lm(formula = O ~ WS + Temp + H, data = pollut)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5887 -1.1686  0.1978  1.9004  4.1544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.60697   13.07154  -1.270    0.215
WS           -0.44620    0.08513  -5.241 1.78e-05
Temp          0.60190    0.11764   5.117 2.47e-05
H             0.09850    0.06316   1.559    0.131

Residual standard error: 2.874 on 26 degrees of freedom
Multiple R-squared:  0.7964,        Adjusted R-squared:  0.7729
F-statistic: 33.89 on 3 and 26 DF,  p-value: 3.904e-09

> newdata = data.frame(WS = 40, Temp = 80, H = 50)
> predict(pollut_step, newdata, interval = "confidence")

      fit      lwr      upr
1 18.6218 16.70852 20.53509

> predict(pollut_step, newdata, interval = "prediction")

      fit      lwr      upr
1 18.6218 12.41146 24.83215

> library(ggfortify)
> autoplot(pollut_step, which = 1:2) + theme_bw()
```



8. Adapted from Lab 4B Who has diabetes?

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict whether or not a patient has diabetes based on certain clinical measurements. The larger database of patients has been subset such that all observations here are females at least 21 years old of Pima Indian heritage.

The data sets consists of several medical predictor variables and one target variable, 'y' which equals 1 if an individual is diabetic and 0 otherwise. Predictor variables includes the number of pregnancies the patient has had ('npreg'), their BMI, insulin level ('serum'), age, triceps skin fold thickness 'skin', diastolic blood pressure ('bp'), plasma glucose concentration ('glu') and diabetes pedigree function ('ped').

R output is given below to help you answer the following questions.

(a) Write down the fitted stepwise model.

(b) In the stepwise model, increases in which variables lead to higher odds of diabetes?

(c) Predict the log-odds and the probability of a 35 year old woman who has been pregnant twice, with a BMI of 30, blood pressure of 72, glucose of 122 and diabetes pedigree function of 0.5. Compare it to another woman with the same measurements, except a BMI of 40.

(d) Calculate accuracy, sensitivity and specificity to assess the in-sample accuracy of the predictions from the stepwise model.

(e) A decision tree was also estimated to predict diabetes outcome. Predict the outcomes for the two women described earlier using the estimated tree.

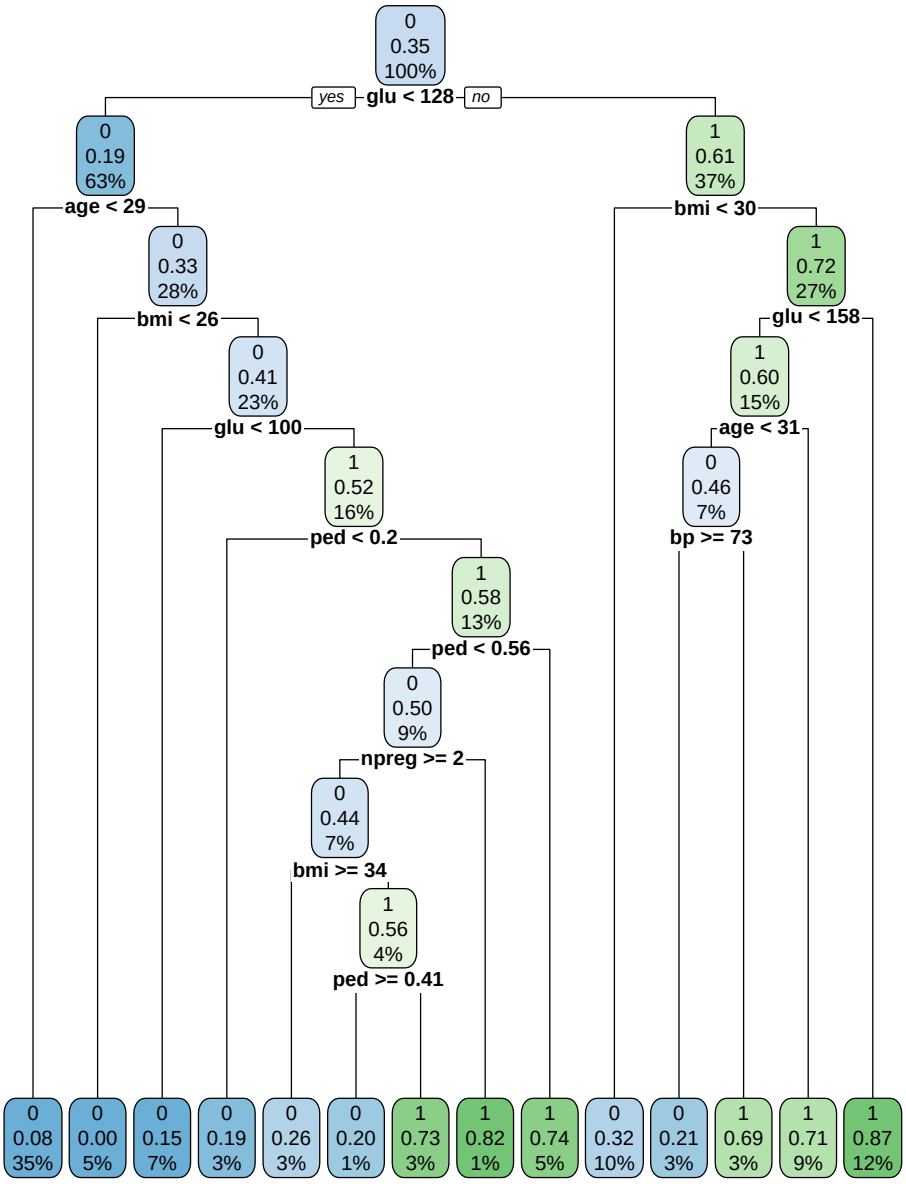
(f) Repeated 5-fold cross validation was performed using the caret package. When comparing between logistic regression, random forest and decision tree, which method do you prefer and why.

(g) Describe one advantage of a random forest relative to a decision tree and one advantage of a decision tree relative to a random forest.

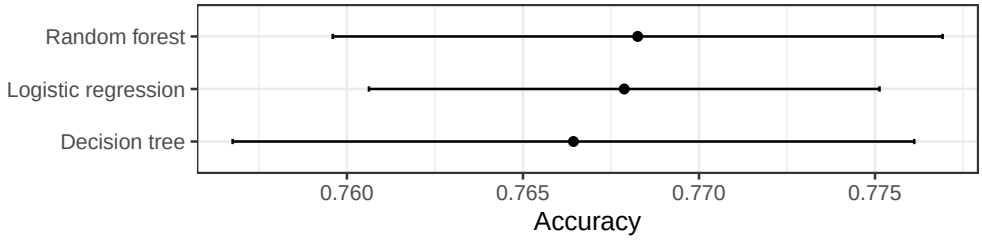
```
> pima = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/pima.csv")
> pima = pima %>%
+   dplyr::select(-serum, -skin) %>%
+   mutate_at(.vars = vars(bmi, bp, glu),
+             .funs = funs(ifelse(. == 0, mean(., na.rm = TRUE), .)))
> glimpse(pima)

Rows: 768
Columns: 7
$ npreg <dbl> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, 5, 7, 0, 7, 1,...
$ glu   <dbl> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125, 110, 168, 13...
$ bp    <dbl> 72.00000, 66.00000, 64.00000, 66.00000, 40.00000, 74.00000,...
$ bmi   <dbl> 33.60000, 26.60000, 23.30000, 28.10000, 43.10000, 25.60000,...
$ ped   <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.248, 0.134, 0.1...
$ age   <dbl> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 34, 57, 59, 51,...
$ y     <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1,...

> fm = glm(y ~ ., data = pima, family = binomial)
> step_model = step(fm, trace = FALSE)
> summary(step_model)

Call:
glm(formula = y ~ npreg + glu + bmi + ped, family = binomial,
    data = pima)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.18863    0.70572 -13.020  < 2e-16
npreg        0.14338    0.02755   5.205 1.94e-07
glu          0.03691    0.00349  10.575  < 2e-16
bmi          0.08871    0.01473   6.024 1.71e-09
ped          0.88252    0.29477   2.994  0.00275

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 716.17  on 763  degrees of freedom
AIC: 726.17

Number of Fisher Scoring iterations: 5
```

```
> new_data = data.frame(age = c(35, 35),
+                       npreg = c(2, 2),
+                       bp = c(72, 72),
+                       bmi = c(30, 40),
+                       glu = c(122, 122),
+                       ped = c(0.5, 0.5))
> predict(step_model, new_data, type = "link")

         1          2
-1.2966884 -0.4096245

> predict(step_model, new_data, type = "response")

        1         2
0.2147229 0.3990022

> library(caret)
> preds = as.factor(round(predict(step_model, type = "response")))
> truth = as.factor(pima$y)
> table(preds, truth)

     truth
preds   0   1
    0 443 118
    1  57 150

> trctrl = trainControl(method = "repeatedcv", repeats = 10, number = 5)
> logistic_cv = caret::train(factor(y) ~ npreg + glu + bmi + ped,
+       data = pima, method = "glm", family = "binomial", trControl = trctrl)

> library(rpart)
> library(rpart.plot)
> tree = rpart(factor(y) ~ ., data = pima)
> tree_cv = caret::train(
+   factor(y) ~ .,
+   data = pima, method = "glm", family = "binomial", trControl = trctrl)

> rpart.plot(tree, cex = 0.85)
```




```
> rf_cv = caret::train(factor(y) ~ .,
+       data = pima, method = "rf", trControl = trctrl)
> list(`Random forest` = rf_cv, `Decision tree` = tree_cv, `Logistic regression` = logistic_cv) %>%
+   resamples() %>% ggplot() + theme_bw() + labs(y = "Accuracy")
```



9. Describe the process of k-means clustering.

[100 words or less.]

10. What is the purpose of principal component analysis? How can we select the number of principal components we need to retain?

[100 words or less.]

11. Why is an ANOVA post-hoc t-test generally considered preferable to standard two-sample t-test?

[100 words or less.]

12. You're the senior manager at a management consulting company. A junior data analyst on your team has been tasked with building a prediction model for a binary outcome. When you ask them how their model performs they respond with:

"It's awesome, bro! Best model ever. I used all available variables in a logistic regression model. The resubstitution accuracy was pretty much the same as the leave-one-out cross validation accuracy. So I'm done for the day. I'm gonna go play some fussball and grab a kombucha from the fridge, can I get you one bro?"

You remind the junior analyst for the 100th time that you're not their "bro". Internally you curse the HR department for hiring Commerce grads from UNSW.

In the text box below, provide some guidance to the junior analyst about their model selection and evaluation choices. Also suggest some alternative methods that they could use and briefly outline their advantages and disadvantages.

[150 words or less.]