Problem 3 (1 credit)

Load the data set gapminder from the library dslabs using the following code:

dt <- data.table(dslabs::gapminder)

Create a table that contains the infant_mortality for each year starting from 1980 and for each of the following countries: Estonia, Romania, Montenegro and Israel. Write R code using the library data.table to create such a table. The header of such a table is displayed below. Maintain the same column names. (1 point for the correct solution.)

year	Estonia	Israel	Montenegro	Romania
1980	22.4	15.3	NA	34.9
1981	21.6	14.6	NA	33.6
1982	20.9	14.0	NA	32.5
1983	20.2	13.5	NA	31.8
1984	19.6	12.8	23.8	31.6
1985	19.0	12.1	21.8	31.8

```
dt <- data.table(dslabs::gapminder)
dt
dt <- dt[, .("country", "infant_mortality") ]
dt <- dcast(dt, ... ~ country)
dt
```

Problem 4 (2 credits)

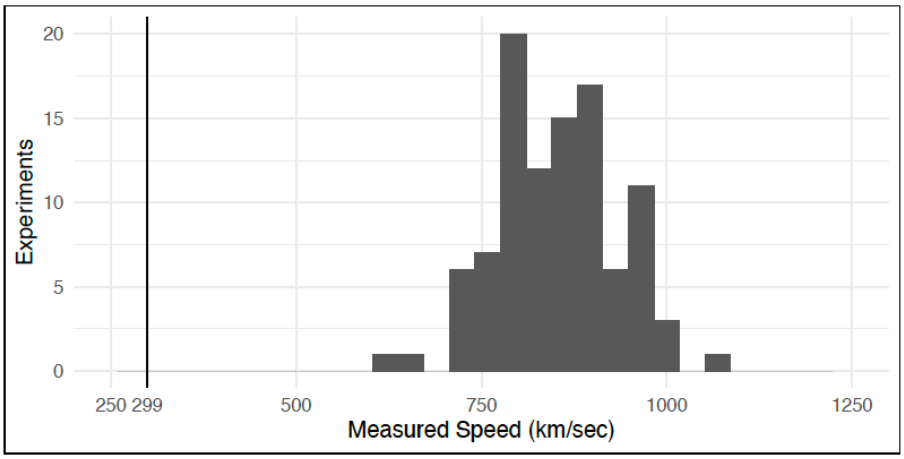
The morley dataset from the datasets package contains 100 measurements done in 1879 on the speed of light. The data consists of five experiments, each consisting of 20 consecutive 'runs'. The response is the speed of light measurement, suitably coded (km/sec, with 299000 subtracted).


```
plot.data <- as.data.table(morley)
head(plot.data)
```



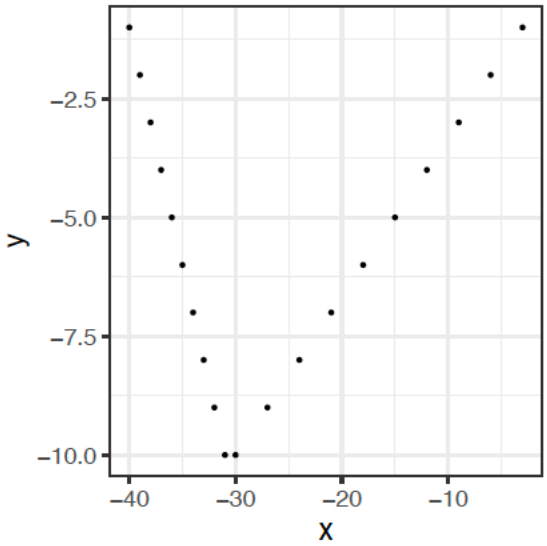
```
##    Expt Run Speed
## 1:    1   1   850
## 2:    1   2   740
## 3:    1   3   900
## 4:    1   4  1070
## 5:    1   5   930
## 6:    1   6   850
```

Write R code that produces the following plot. Matching the exact rendering style (font, font size, choice of color) is not requested. (1 point for the correct data wrangling. 1 point for the correct plot.)



Problem 5 (2 credits)

The plot below displays the numerical association between x and y.

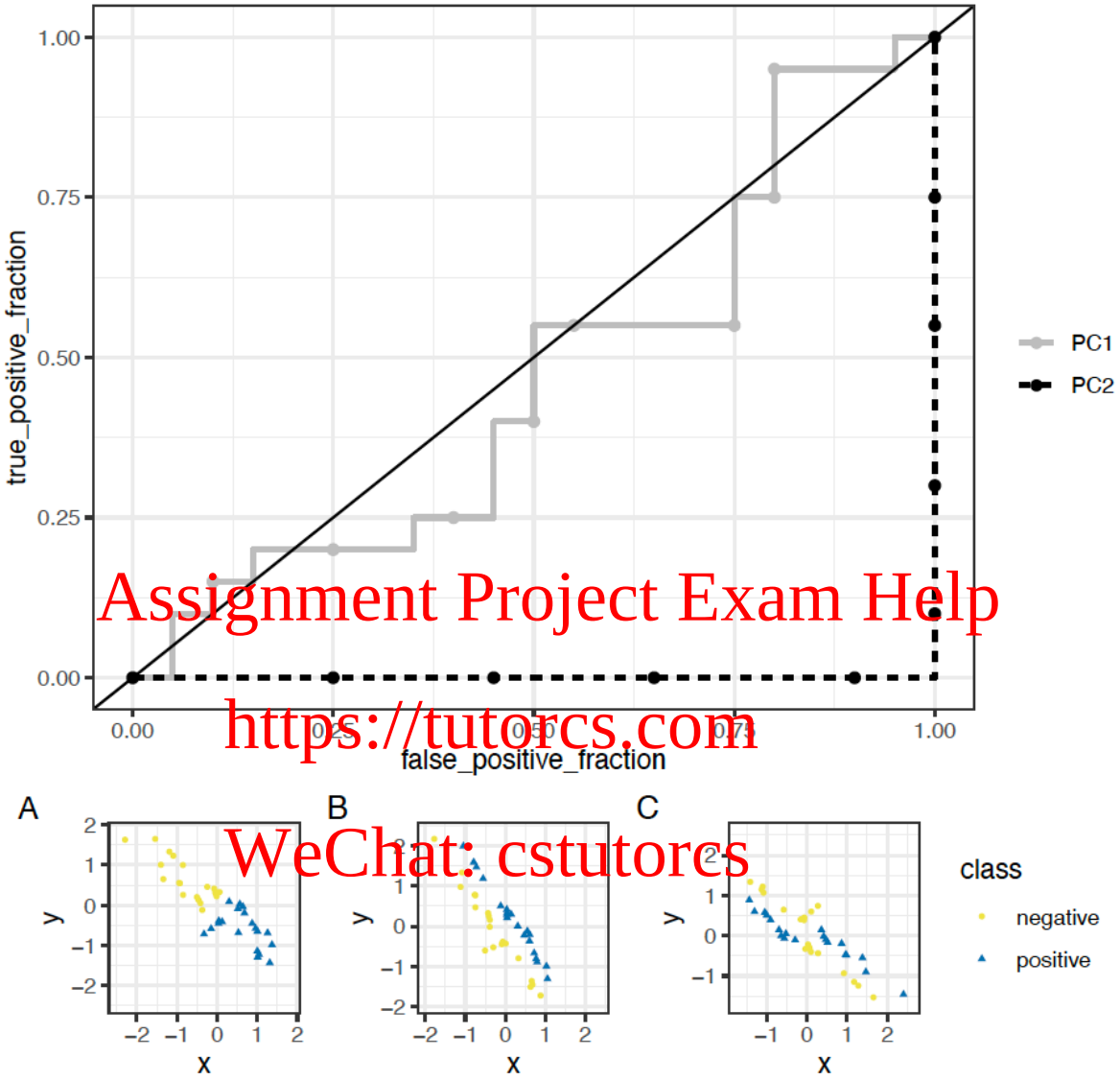


One of the following options i-v is correct for each correlation coefficient c:

i) c = -1
ii) -1 < c < 0
iii) c = 0
iv) 0 < c < 1
v) c = 1

a) Indicate the correct option for the Pearson correlation coefficient and justify your answer.
b) Indicate the correct option for the Spearman correlation coefficient and justify your answer.
(For each a) and b) 1 point for the correct answer with a correct justification.)

Problem 6 (1 credit)



The upper plot shows ROC curves for discriminating points of a positive class (blue triangles) from points of a negative class (yellow circles) using the projections on PC1 (full grey line) and PC2 (dashed black line) as scores. Which dataset, A, B or C was the PCA computed on? Justify your answer using the definition of PCA and ROC curve. (1 point for the correct answer and 1 point the correct justification.)

Problem 7 (4 credits)

Consider the shuffled version of the 'tissue_gene_expression' dataset generated with the following code:

```
set.seed(13)

library(dslabs)
dt <- as.data.table(tissue_gene_expression$x)
dt[, y := tissue_gene_expression$y == "liver" ]
dt <- dt[sample(1:nrow(dt))] # this reshuffling prevents convergence errors of logistic regression
```

Define a training set containing only the first 113 observations of the shuffled dataset and a test set containing the remaining observations. Fit a logistic regression model to predict the response variable gene expression in liver (variable y =="liver") on the training set with only the explanatory variable PLEKHJ1.

a) Evaluate the model by computing the area under the curve (AUC) of ROC on the above defined test set using the functions calc_auc() and geom_roc() from the library plotROC. Provide R code and explicitly state the computed value rounded to two significant digits. (1 point for code properly computing predictions on test set. 1 point for the correctly computed and stated AUC.)

b) Wrap the section of your code from (a) that computes the AUC on the defined test set into a function with the following structure:


```
get_AUC <- function(dt_test) {
AUC_on_test <- # your code computing AUC on test set here
return(AUC_on_test) }
```


Using this function, obtain a 80 % equi-tailed bootstrap confidence interval for the computed AUC on the above defined test set by resampling the computation of AUC. Run 99 bootstrap iterations. Provide R code and explicitly state the lower and upper bound of the interval rounded to two significant digits. (1 point for code that implements the bootstrap relying on a get_AUC() function (itself be correct or not). 1 point for correct bootstrap interval bounds.)

Problem 8 (2 credits)

Consider the dataset olive. Do not make any assumption of normality. Assuming a significance level of 0.05, which statistical test that we studied do you suggest to test the association between the variable palmitoleic and the variable linolenic >=0.33? Justify, in one sentence, why you used this statistical test. Provide the P-value rounded to two significant digits and state whether you reject the null hypothesis. (1 point for the correctly justified statistical method. If so, 1 further point for the R code, the correct numerical answer and a correct answer to the question whether the null should be rejected.)

Problem 9 (2 credits)

Alice and Bianca would like to predict the variable Petal.Length from the dataset iris using linear regression. In this question, assume that the assumptions of linear regression hold in every regression performed. Load the dataset using the command: data("iris").

a) Alice has proposed a model using only Sepal.Width as a predictor. Bianca, argues that we also need to include Sepal.Length. How much more of the variance is explained by Bianca's model? Provide R code and the added fraction of the variance rounded to 2 significant digits. (1 point for the R code and the correct value).
b) Alice disagrees, arguing that this additional variable does not significantly improve the model. Settle the debate with an appropriate test. Provide R code as well as the P-value rounded to 2 significant digits. (1 point for the R code and the correct value).

Problem 10 (2 credits)

Consider the R code below that defines the variables A,B,C and D:

```
n <- 200
sigma <- function(z){ 1 / (1 + exp(-z)) }
C <- runif(n)
A <- rnorm(n, mean=C)
D <- C + rnorm(n)
B <- rbinom(n, size=1+round(exp(C)), prob=0.5)
```


a) List all the directed edges (e.g. A->B) of the causal diagram between the variables A, B, and C and no other potential variables. (1 point for the exact list)

b) List all the pairwise associations (statistical dependencies) as undirected edges (e.g. A-B) between the variables A, B, and C and no other potential variables. (1 point for the exact list and justification naming the relevant causal diagram scenario).

Problem 11 (4 credits)

The 10 turbines of a power plant must be monitored today. If a turbine is working properly, its temperature follows a normal distribution, with a mean of 787 Kelvin and a standard deviation of 15 Kelvin. Today your 10 employees measure the temperature of his or her assigned turbine independently and obtain the following measurements:

```
temps <- c(751.78, 750.92, 790.82, 752.96, 784.59, 820.7, 835.03, 787.53, 827.67, 793.32)
```

a) In a first approach, you request each employee to turn off his or her turbine if the null hypothesis that the turbine is working properly can be rejected at the 0.025 significance level. How many employees will turn off his or her turbine? Provide R code to determine this number and state it in a plain English sentence. Justify your answer. (1 point for the justification and 1 point for R code and the correct number of defective turbines)
b) In a second approach, you want to avoid unnecessarily shutting off turbines, as it is very expensive. You want you to ensure that the proportion of properly working turbines among those turned off is at most 2.5 percent. Under this criterion, how many turbines must be turned off? Provide R code to determine this number and state it in a plain English sentence. Justify your answer. (1 point for the justification and 1 point for R code and the correct number of defective turbines).

Problem 12 (2 credits)

Alice would like to fit a linear regression model to predict the variable oleic from the dataset olive using the variables arachidic, linolenic and palmitic as predictors. Provide R code that implements a linear regression in a 5-fold cross validation. Compute the R-squared on each validation fold and return the minimal value of these computed values.

Load the dataset with the following command:
dt <- data.table((na.omit(dslabs::olive)))

For reproducibility, do not reorder the rows of the dataset and define the fold index of each sample with the following command:

dt[, fold := cut(1:.N, breaks=5, labels=F)]

In this question it is not asked to check whether the assumptions of the linear regression reasonably hold for the data. You are only allowed to use the library data.table and R built-in commands to solve this question. (1 point for the correct R code. 1 point for the correctly computed minimal value of the R-squared over all validation folds in a plain English sentence.)

Problem 13 (2 credits)

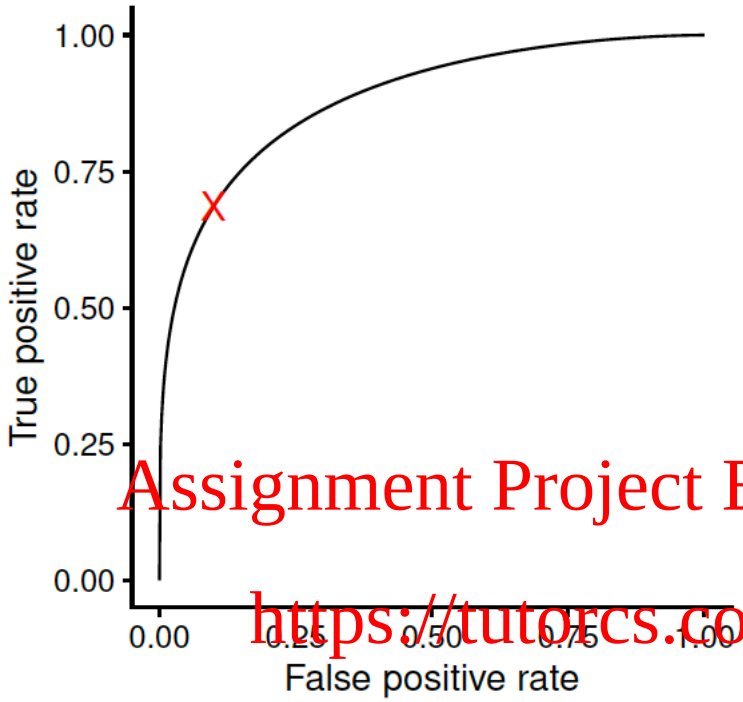
A clinical trial was conducted in Datavizland to test the efficacy of a new vaccine in preventing viral infection. The summary of the results can be constructed using the following code:


```
tbl <- matrix(c(75,94,126,28),nrow=2,ncol=2)
row.names(tbl) <- c("Not Infected","Infected")
colnames(tbl) <- c("Not Vaccinated","Vaccinated")
```


Assuming a significance level of 0.1, and using an appropriate statistical test and a relevant alternative, do you reject the null hypothesis that there is no association between being vaccinated and avoiding infection? Justify, in one sentence, why you used this test and alternative. Provide the P-value rounded to two significant digits and state whether you reject the null hypothesis. (1 point for the correctly justified statistical method. If so, 1 further point for the R code, the correct numerical answer and a correct answer to the question whether the null should be rejected.)

Problem 14 (2 credits)

A classifier is trained on a dataset to discriminate individuals affected by a disease versus healthy individuals. This dataset consists of 1,000 diseased cases (positives) and 1,000 healthy controls (negatives). We obtain the following cross-validation ROC curve (AUC= 0.88), where the suggested classifier cutoff is marked by a cross (FPR = 0.1 , TPR= 0.70).



a) What is the value of the AUC, if the ROC curve is computed on a very large population where the odds for the disease are 1:100? Justify your answer providing relevant equations and derivation steps when need be. Provide the result with two significant digits. R code is not required but may be used. (1 point for the correct answer and justification.)

b) What is the value of the precision at the same suggested classifier cutoff if the ROC curve is computed on a very large population where the odds for the disease are of 1:100? Justify your answer providing relevant equations and derivation steps when need be. Provide the result with one significant digit. R code is not required but may be used. (1 point for the correct answer and justification.)