Candidate Number:

ISSU0053 Data Science and Big Data Analytics

UCL Summer School 2019

Assessment II **Computer Practical under examination conditions (50%)**

Timing **Friday 9th August**

10:00 – 12:30

Examination conditions:

- **The session is open book, so you can consult your notes, textbooks and programming websites as you work.**
- **You must work on the UCL desktop.**
- **All work submitted must be your own. All forms of communication and messaging with other students is strictly prohibited, and violations will be dealt with in accordance with UCL policy.**
- **At the end of the assessment you will have a short time to collate the files you have produced and upload them for marking.**

Notes:

Marks for questions are indicative, and a grading curve may be applied to generate a final grade.

All code used to complete the tasks must be submitted in R script files or R notebook files in Rmd format. Partial marks will be awarded for code sections that have been completed but are non-functional.

You should work in one file per section, so at the end of the exam upload either e.g.

sectionA.R, sectionB.R, sectionC.R sectionD.R, sectionE.R
or sectionA.Rmd, sectionB.Rmd, sectionC.Rmd, sectionD.Rmd, sectionE.Rmd

It is recommended you use comments / notes in your files to indicate which part of the question you are attempting, e.g.

```
# A1
plot(…)

# A2
```

```
summary(boston)
```

Section (A)
Start a new R file for your answer to this question.

The file ***sales.csv*** contains data on the sales of items in a university campus cafe that opens on Mondays to Fridays and is closed at the weekend.

Write R code to:

A1. Load the contents of the file into a data frame.

(1 mark)

A2. Display the names of the columns that have been loaded.

(1 mark)

A3. Edit the data frame so that column ***windspeed*** is renamed to ***wind***

(1 mark)

A4. Count the number of rows in the data frame.

(1 mark)

A5. Drop all rows with NA values and calculate the number of rows removed.

(2 marks)

A6. Edit the data frame to remove the ***date*** column.

(1 mark)

A7. The column ***staff*** records who was the person working in the shop using the following mapping:

1	***2***	***3***	***4***
Harry	***Sara***	***Tom***	***Kate***

Adjust the data frame so that ***staff*** uses the names as given above in a factor type column.

(3 marks)

A8. Add a column ***food*** that totals the number of pizza, pasta, and wrap sales each day.

(2 marks)

(12 marks)

Section (B)
Start a new R file for your answer to this question.

Run the command below to load the following data file into R.

```
load("section_b.Rda")
```

This contains the data frame: ***sales***. This contains the modified version of the data frame you worked on in Section (A) without the ***food*** column.
Write R code to find:

B1. The total number of soda sales made over all days recorded in the dataset.
(1 mark)

B2. The highest number of wraps sold in a single day.
(1 mark)

B3. The average value of ***total_sales*** on the days when Harry was staffing the shop .
(1 mark)

B4. The average value of ***total_sales*** on the days when Sara was staffing the shop .
(1 mark)

B5. Perform a t-test to test the hypothesis that the average ***total_sales*** achieved is different on the days Harry works in the shop, compared to the days Sara works in the shop.
(1 mark)

You should see the following result.

```
Welch Two Sample t-test

data:  ...
t = 1.9694, df = 24.574, p-value = 0.06028
alternative hypothesis: ...
95 percent confidence interval:
 -1.281428 56.170716
sample estimates:
mean of x mean of y
 163.3293  135.8846
```

B6. Interpret the result against the hypothesis based on a significance level of 0.05.
(1 mark)

B7. Write R code to examine the dataset and display the maximum and minimum humility levels over the recorded period.

(1 mark)

B8. Make the following fits:

model 1. predict sales of coffee using temperature
model 2. predict sales of coffee using humidity
model 3. predict sales of coffee using wind speed

(2 marks)

B9. Display the fits using ***screenreg***.

(1 mark)

```
============================================
               Model 1     Model 2    Model 3
--------------------------------------------
(Intercept)   45.77 ***  -1.66       28.63 ***
              (3.76)     (5.55)      (5.36)
temp          -0.59 ***
              (0.09)
humidity                  0.40 ***
                         (0.09)
wind                                 -0.39
                                     (0.31)
--------------------------------------------
R^2            0.51       0.32        0.04
Adj. R^2       0.50       0.31        0.01
Num. obs.     43         43          43
RMSE           7.46       8.80       10.50
============================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

B10. Explain (with reference to the p-values) how we can interpret the value and significance of the coefficients associated with temperature, humidity and windspeed.

(3 marks)

B11. Use the fit of model 2 to predict of the number of coffees sold when:

i) the humidity level is 30 ii) the humidity level is 80.

(2 marks)

(15 marks)

Section (C)
Start a new R file for your answer to this question.

In this question we will try to predict ***pizza*** sales based on the other column values.
Start by loading the dataframe ***sales*** using command:

```
load("section_b.Rda")
```

C1. Write code to display the frequency table for daily pizza sales.

(1 mark)

C2. Build a linear regression model to predict pizza sales from the other columns using forward stepwise selection until an optimal model is reached.

(3 marks)

If you are unable to complete this question. You can generate the resulting fit using the code below (also found in the ***code_example.R*** file):

```
lm_step = lm(pizza ~ weekday + humidity + staff + dessert +
total_sales +  soda, data = sales)
```

C3. Examine the diagnostic plots associated with this model.
Why is row 18 highlighted in the final plot?

(1 mark)

C4. Write an R command that displays the data associated with this row.

(1 mark)

C5. Do you feel this row should be considered an outlier to the rest of the dataset?
Explain your answer.

(1 mark)

C6. Write R code that uses the fitted residuals to calculate the RSS value for the fit.
(You should find this to be 81.712).

(1 mark)

C7. Calculate the estimated RMSE of the fit using

$$estimatedRMSE = \sqrt{\frac{RSS}{43-11}}$$

(1 mark)

C8. The number 43 in the formula above refers to the number of rows.
What does the other number refer to?

(1 mark)

C9. Which of the terms MSE, RMSE, TSS and RSS is of most use to someone considering the accuracy of a prediction produced by the model?

(1 mark)

C10. Use R to build a simple model that predicts pizza sales using ***`weekday`*** as a single predictor.

(1 mark)

This should generate a table of coefficients like:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.1667     0.8058   2.689  0.01058 *
weekdayMonday      1.3889     1.0402   1.335  0.18977
weekdayThursday    3.7083     1.0659   3.479  0.00128 **
weekdayTuesday     4.5333     1.0192   4.448 7.32e-05 ***
weekdayWednesday   2.8333     1.0192   2.780  0.00841 **
```

C11. Examine the coefficents. How can we interpret the meaning of these values?

(2 marks)

C12. Use R to compare the performance of this model with the larger model from stepwise selection using an ANOVA test.

(1 mark)

This should produce the following table:

```
Analysis of Variance Table

Model 1: pizza ~ weekday
Model 2: pizza ~ weekday + humidity + staff + dessert
                     + total_sales + soda
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     38 148.031
2     32  81.712  6    66.318 4.3286 0.002635 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

C13. Interpret the result of the ANOVA test with reference to the coefficients of the additional predictors included in the larger model.

(1 mark)

C14. Explain why even through the performance of the larger model is better it is of no practical purpose in predicting pizza sales.

(1 mark)

C15. Why do we use cross validation methods to compare models when we already have estimates like R^2, Adjusted-R^2, and estimated RMSE to measure performance?

(2 marks)

C16. When performing k-fold validation using ***`cvFit/trainControl`*** we can provide an argument R/repeats.
Explain what this argument does, and why it is useful to make use of it.

(2 marks)

C17. When performing k-fold on this data set what is the maximum number of folds we could use?

(1 mark)

C18. Perform k-fold validation on the following models with R=100 and 10 folds. Include a suitable command to ensure that your analysis is reproducible.

(4 marks)

model 1. predict number of pizza sales based on weekday
model 2. predict number of pizza sales based on weekday, temperature, humidity and windspeed

You should expect to obtain results that are similar (but not identical) to one of the tables below:

Predictors	**RMSE**	**Standard Error**
day	2.13774	0.1219257
day and weather columns	2.110702	0.1262512

Predictors	**Residual standard error:**	**Adj R^2**
day	1.974	0.3434
day and weather columns	1.855	0.4197

C19. Which of these models is best to use in this case? Explain your reasoning.

(1 mark)

(27 marks)

Section (D)
Start a new R file for your answer to this question.

Load in the diabetes data set ***medical_data*** using the command.

```
load("section_d.Rda")
```

D1. Make two histogram plots showing the distributions of the entries in columns ***glucagon***, and ***log_glucagon***

(2 marks)

D2. Based on these plots would ***glucagon*** or ***log_glucagon*** be better to use as a predictor in a logistic regression/LDA classification analysis? Explain your answer.

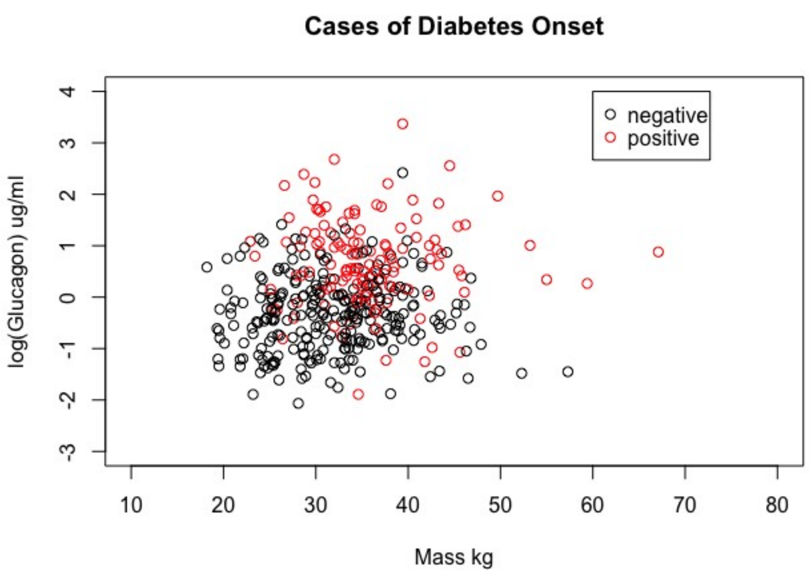
(2 marks)

D3. Write the R command that could be used to create column ***log_glucagon*** based on the values in the ***glucagon*** column.

(1 mark)

D4. Use the data in ***medical_data*** to create a plot similar to the one shown below:

(6 marks)



allocation of marks for:

1 mark – displaying the correct data
1 mark – matching x and y axis limits
1 mark – matching x and y axis labels
1 mark – matching title
1 mark – matching legend
1 mark – colouring points according to diabetes category (red: positive & black: negative)

You may use ***`ggplot`*** if you wish.

PLEASE CHECK RESOURCE FOLDER FOR COLOUR VERSION OF PLOT!

D5. Look at the plot. Do you think it will be possible to use log_glucagon level and mass to predict whether a person will be in the positive group, i.e. developed diabetes? Explain your answer.

(2 marks)

In order to run both logistic regression and K-nearest neighbours methods we will create columns that store rescaled predictors for ***log_glucagon*** and ***mass***.
This can be done by applying the scale function:

```
medical_data$log_glucagon_rs <- scale(medical_data$log_glucagon)
medical_data$mass_rs <- scale(medical_data$mass)
```

D6 . Write R code that displays the mean and standard deviation of the rescaled columns, and comment on how the columns are now scaled.

(2 marks)

To run our analysis we will create a dataframes ***train_data*** containing a random sample containing 60% of the full dataset. The rest of the data will be used as a validation set and stored in data frame ***test_data***.

D7. Write R commands that shows how to create these data frames from ***medical_data***.

(3 marks)

(18 marks)

Section (E)
Start a new R file for your answer to this question.

In order to ensure consistency load in the ***train_data*** and ***test_data*** data frames that have been generated for you:

```
load("section_e.Rda")
```

E1. Perform a logistic regression to predict the diabetes class based on the ***log_glucagon_rs*** and ***mass_rs*** values.

(2 marks)

E2. Use the resulting model to make predictions for the test data set, predicting diabetes if the model predicts more than a 50% or 0.5 probability for a positive diagnosis.

(2 marks)

E3. Create a confusion matrix in the following format. You should find the following result:

(2 marks)

```
              actual
    predicted neg pos
          neg  97  24
          pos   7  30
```

E4. How many total cases of positive diabetes onset were there in the test dataset?

(1 mark)

E5. How many of these cases of diabetes onset were correctly predicted by the model?

(1 mark)

E6. Hence determine the true positive rate TPR:

$$TPR = \frac{positive cases correctly identified}{total actual positive cases}$$

(1 mark)

E7. Make a new set of predictions, placing people in the positive category if they are predicted to have more than a 30% or 0.3 probability of being in the group of people developing diabetes.

(1 mark)

```
          actual
predicted neg pos
      neg  78   9
      pos  25  45
```

E8. Show that this increases the true positive rate (TPR).

(1 mark)

E9. Suggest a reason why this could be advantageous.

(1 mark)

E10. In what respect has making this change decreased our measured performance?

(1 mark)

We will now try to classify the cases using k-nearest neighbours method.

E11. From ***train_data*** and ***test_data*** create the following:

train_data_x a data frame containing rescaled predictors for glucose and mass

train_data_y a column vector containing diabetes class.

test_data_x a data frame containing rescaled predictors for glucose and mass

test_data_y a column vector containing diabetes class.

(4 marks)

Note:
if you are using KNN(): If you cannot complete E11. you can load these objects from the ***section_e_knn.Rda*** *file.*
If you are using caret::train(): please still complete this step but proceed below as usual

E12. Apply the method of KNN to predict the class of the test data set with maximum 9 nearest neighbours.

(3 marks)

Note if you are using KNN(). The ***knn*** function requires the ***train_data_y*** to be 1D object. Depending on the way you create ***train_data_y*** and ***test_data_y*** and you may end up with a 2D data frame that consists of a single column. In this case you can convert them using commands like:

```
train_data_y <- unlist(train_data_y).
```

E13. Calculate the misclassification rate.

(1 mark)

E14. The following code creates a sequence of 30 numbers running from 2 to 200:

```
K_values = c( seq(2,12,2), seq(15,55,5), seq(60,200,10) )
```

Write code to test the KNN method using these values for K, and plot
- if you use KNN(): the resulting misclassification rate against K.
- if you use caret::train(): the resulting accuracy against K.

(4 marks)

E15. Code a method to repeat the KNN analysis resampling the data into train and test groups several times (20 repeats) in order to generate a measure of averaged performance.

(3 marks)

(32 Marks)