PAPER CODE NO. **COMP336**	EXAMINER: DEPARTMENT:	Dominik Wojtczak Computer Science	Tel. No. dkw@liverpool.ac.uk

UNIVERSITY OF LIVERPOOL

FIRST SEMESTER EXAMINATIONS 2023/24

Big Data Analytics

TIME ALLOWED : TWO Hours

INSTRUCTIONS TO CANDIDATES

Answer **ALL** questions.

The exam consists of 4 questions, each worth 25 marks.

The numbers in the right hand margin represent an approximate guide to the marks available for that part of a question.

Total marks available are 100.

Calculators are permitted.

1. **(a)** A company are experimenting with a chemical process to try and increase its yield. The table below contains the results of their first few experiments into the amount of power they use to heat their vessel (x) and the yield of product produced (y).

Power x	Yield y
1.0	18.5
2.2	-10
3.5	22.0
4.0	223.5
5.0	26.0
6.0	28.5

Comment on whether any of the values are unusual. How would you prepare this data to be used in a statistical model? **[4 marks]**

(b) Give two ways that this data can be investigated before modelling. **[2 marks]**

(c) They want to model the data using linear regression. State two assumptions that are made about the model errors when using a least-squares linear regression model. **[2 marks]**

(d) State the quantity that is minimised to obtain the parameters of a least-squares model. **[2 marks]**

(e) After some more experiments they have collected 10 data points. Given the sums of variables in the table below, calculate estimates for the parameters of a simple linear regression model for y based on x.

n	10
$\sum_i x_i$	56
$\sum_i x^2$	397
$\sum_i y_1$	271
$\sum_i x_i y_1$	1688

[10 marks]

(f) A linear regression model has a total error (measured using the quantity in **1.(d)**) of 88. When a new feature is added and the model re-calculated, the error is reduced to 82. Is the second model necessarily better? How could you check that the new model is an improvement? **[3 marks]**

(g) If you now want to predict the probability of a single event using the same input features, how might the linear regression model be adapted? **[2 marks]**

2. (a) Suppose we have a cluster with 3 data nodes: node DN1, DN2 and DN3. We want to distribute 3 files from a local file system to this HDFS cluster with replication factor 2. File 1 is divided into two blocks: B1 and B2, File 2 into two blocks: B3 and B4, and File 3 into two blocks: B5 and B6. Your task is to distribute all blocks across our data nodes DN1, DN2 and DN3 as evenly as possible. **[5 marks]**

(b) Consider the following dataset: $\mathbf{x} = \{1, 7, 7, 7, 8, 36\}$.

i. Calculate the mean and variance of this dataset. **[7 marks]**

ii. Do you think that the mean and variance are representative of the location and spread respectively? Justify your answer. **[3 marks]**

iii. What other summary statistics could you use to describe the location and spread? **[4 marks]**

(c) What is more likely when tossing a fair coin 4 times: there will be 3–1 split (i.e., 3 tails and 1 head, or 3 heads and 1 tail) or there will be 2–2 split (i.e., 2 heads and 2 tails)? Justify your answer and calculate the probability of these events. **[6 marks]**

3. Consider the following dataset, which consists of a set of records with the format:

(month, day of the month, number of bikes rented, cost)

Write down PySpark code that just uses the standard RDDs' actions and transformations to output what each of the questions **(a), (b), (c), (d), (e)** below asks for. (We are only interested in the code, not its output.) You should assume that everything is already setup, and the data is loaded into variable `input` which consists of 4-tuples (i.e., not a DataFrame). In other words, if x is a single row, then $x[0]$ is the month, $x[1]$ is day of the month, $x[2]$ is the number of bikes hired, and $x[3]$ is the cost of hiring a bike. Please note that there can be multiple different costs of hiring a bike reported on the same date. In each of the solutions you have to use `reduceByKey` transformation. The reduceByKey(F) transformation combines values with the same key using a specified function F and reduces them to a single value. In particular, you can make use of the standard function `add`, which is essentially `lambda x, y:  x + y`.

(a) For each month output the the total hire cost (i.e., the sum over all days that month of the number of bikes hired times its cost).

[5 marks]

(b) Output the month for which the cost of a bike hire was the highest.

[5 marks]

(c) For each (month, day) pair output the total number of bikes hired that day.

[5 marks]

(d) Output the (month,day) pair for which the number of hired bikes is the highest.

[5 marks]

(e) Output the average cost of hiring a bike across the whole dataset (i.e., the sum of the number of bikes hired times its cost divided by the total number of bikes hired).

[5 marks]

4. This question concerns the DBSCAN clustering algorithm.

(a) DBSCAN has two parameters: Eps and MinPts. Please define these two parameters. **[4 marks]**

(b) Plot the following dataset of (x, y) coordinates: (0, 0), (1, 0), (2, 0), (5, 3), (5, 4), (5, 5), (0, 5). **[1 mark]**

(c) For an Eps value of ϵ = 1.2, add to your diagram the ϵ-neighbourhoods for all of the seven points. **[2 marks]**

(d) For a MinPts value of 2, label all the points as Core, Border or Noise. **[7 marks]**

(e) Is the point (0, 0) directly density-reachable from (2, 0)? Explain your answer. **[5 marks]**

(f) Keeping MinPts = 2, what is the minimum Eps for which the dataset forms a single cluster? Justify your answer. **[3 marks]**

(g) Using the Eps value that you have just identified, and keeping MinPts = 2, which points are now core points? Justify your answer. **[3 marks]**