UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER 2

WQD7003 : Data Analytics

June 2019 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

(This question paper consists of 5 questions on 4 printed pages)

1. a) What is the difference between data analytics and data analysis? (3 marks)

   b) What is an outlier? (1 mark)

   d) State 2 reasons on why is data dirty? (3 marks)

   e) How missing data is handled? (3 marks)

2. a) Under what conditions might a median be a better measure of the center of your data set than the mean? (3 marks)

   b) What are the criteria of effective visualization? (3 marks)

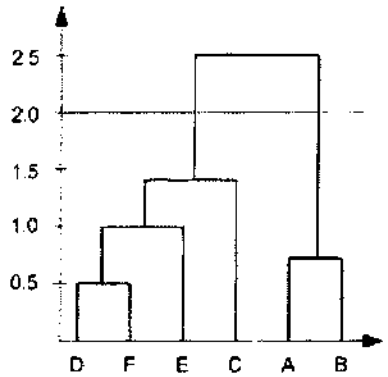
   c) What is data ink ratio? (1 marks)

   d) You need to plot the split between the Republican and Democratic vote for every presidential election since 1990. What chart will you choose? Explain your choice. (3 marks)

3. a) Cross-fertilizing a red and a white flower produces red flowers 25% of the time. Now we cross-fertilize five pairs of red and white flowers and produce five offspring. What is the probability that there are no red flower plants in the five offspring? (3 marks)

   b) What is the minimum no. of variables/ features required to perform clustering? (1 mark)

   c) In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?



(3 marks)

d) Assume you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}

What will be the Manhattan distance for observation (9,9) from cluster centroid C1 (3 marks)

4. a) Explain briefly the 80/20 rule, and its importance in model validation. (3 marks)

b) Explain briefly on cross-validation. (2 marks)

c) What is the difference between supervised and unsupervised machine learning? (2 marks)

d) When should you use classification over regression? (3 marks)

5. a) Following are the results obtained for predicting fraud in credit card transaction.

		Actual	
		Fraud	Not Fraud
Predicted	Fraud	3	97
	Not Fraud	0	0

Calculate the following:-

(i) Precision
(ii) Recall
(iii) F-Score

(3 marks)

b) What will be the output for the following codes:-

(i)
```
a=[1,2,3,4,5,6,7,8,9]
print(a[:2])
```
(1 mark)

(ii)
```
def sum_list(items):
    sum_numbers = 0
    for x in items:
        sum_numbers += x
    return sum_numbers
print(sum_list([1,2,-8]))
```
(1 mark)

(iii)
```
f = lambda x, y : x + y
            f(2,3)
```

(1 mark)

(iv)
```
a = ['Orange','Banana','Apple']
b = ['Banana', 'Orange','Durian']

def union_of_lists(seta,setb):
    return set().union(seta,setb)
```

(2 marks)

(v)
```
import pandas as pd
 import numpy as np
 exam_data  = {'name': ['Anastasia', 'Dima',
'Katherine'],         'score': [12.5, 9, np.nan,],
 'attempts': [1, 3, 2,],
 'qualify': ['yes', 'no', 'yes']}
 labels = ['a', 'b', 'c']
 df = pd.DataFrame(exam_data , index=labels)
 print(df[(df['attempts'] < 3) & (df['score'] > 10)])
```

(2 marks)

END