UNIVERSITY OF MALAYA

EXAMINATION FOR THE MASTER OF DATA SCIENCE

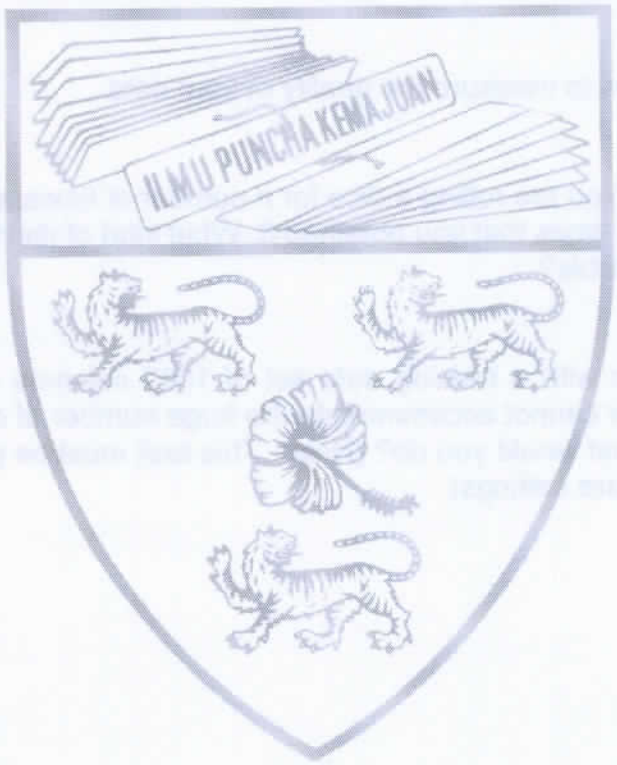
ACADEMIC SESSION 2017/2018 : SEMESTER I

WQD7003: Data Analytics

Jan 2018 Time: 2 hours

INSTRUCTIONS TO CANDIDATES :

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



(This question paper consists of 5 questions on 6 printed pages)

1.

a) List 2 responsibilities of a data scientist.

(2 marks)

b) Differentiate briefly the different types of data analytics

(4 marks)

c) State a methodology that can be used for data analytics project.

(2 marks)

d) Describe a circumstance when median need to be used rather than mean?

(2 marks)

2.

a) State two ways to measure the quality of your data.

(2 marks)

b) Suppose that you are rolling a dice for n number of times and keeping track of the number of times that you obtained 6. What kind of distribution does this scenario resemble?

(1 mark)

c) You are given with a training data set of 1000 columns and 1 million rows. Your computer cannot accommodate the huge number of data due to memory constraint. What would you do? (Note:- This task must be performed with your current hardware settings)

(3 marks)

d) Is the visualizationn shown in Figure 1 effective? What are the criteria(s) that you used in order to evaluate this visualization.

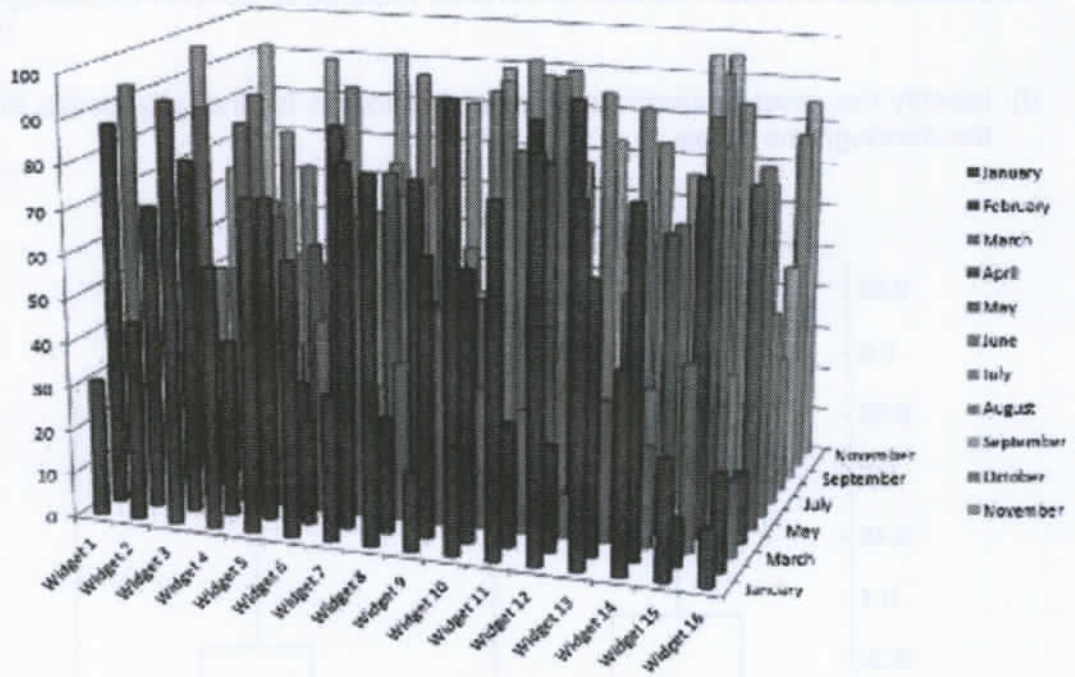


Figure 1: Sales of Widgets Over Time in 3D

(4 marks)

3.

a) Differentiate between continuous and discrete data.

(2 marks)

b) How do you screen for outliers and what would you do to identify it?

(3 marks)

c) The dataset that you are working on contains a lot of noisy data. Describe how would you handle the noisy data?

(4 marks)

d) List 2 techniques for data normalization.

(1 mark)

4.

a)

To test linear relationship of y(dependent) and x(independent) of continuous variables, which of the plot is best suited?

(1 mark)

b) What will happen if we train a linear regression model with less number of data?

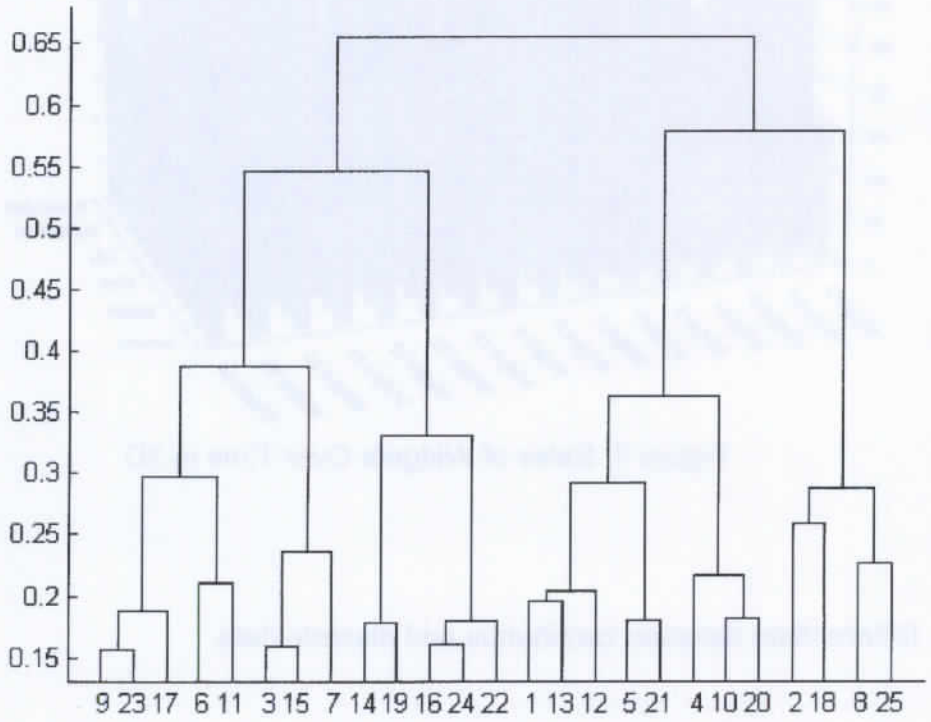
(1 mark)

c) What is the minimum number of variable required to perform clustering?

(1 mark)

d) Identify the most appropriate number of clusters for the data points shown in the dendograms below.

(1 mark)



e) 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

(2 marks)

f) Following are the results observed for clustering 6000 data points into 3 clusters: A, B and C:
Calculate the following:-

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

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

(4 marks)

5.

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

(i)
```
list=[1,2,3,4]
print(list[2:])
```
(1 mark)

(ii)
```
a={"a":1,"b":2,"c":3}
b=dict(zip(a.values(),a.keys()))
b
```
(1 mark)

(iii)
```
for i in range(10):
    if i == 5:
        break
    else:
        print(i)
```
(1 mark)

(iv)
```
f = lambda x, y : x + y
f(2,3)
```
(1 mark)

(v)
```
purchase_1 = pd.Series({'Name': 'Chris',
                        'Item': 'Dog Food',
                        'Cost': 22.50})
purchase_2 = pd.Series({'Name': 'Kevyn',
                        'Item': 'Kitty Litter',
```

```
                    'Cost': 2.50})
purchase_3 = pd.Series({'Name': 'Vinod',
                    'Item': 'Bird Seed',
                    'Cost': 5.00})
df = pd.DataFrame([purchase_1, purchase_2, purchase_3],
  index=['Store 1', 'Store 2', 'Store 2'])

df.loc['Store 2']
```

(2 marks)

b) Explain what the following codes will do:-

(i) df["Score"].fillna(df["Score"].mean(), inplace=True)
df

(1 mark)

(ii) np.var(roll_list_simulated)

(1 mark)

(iii) roll_list_simulated = [ ]

```
for j in range(1000):
        roll1,roll2 = rolltwo()
        roll_total = roll1+roll2
        roll_list_simulated.append(roll_total)

plt.hist(roll_list_simulated, np.arange(1,13)+0.5, normed=True)
plt.xticks(list(range(2,13)))
plt.xlabel('Score')
plt.ylabel('Number of occurences of roll')
plt.title('Histogram of 1000 rolls of 2 dice')
```

(2 marks)

END