NATIONAL UNIVERSITY OF SINGAPORE

BT1101 – INTRODUCTION TO BUSINESS ANALYTICS

(Semester 2: AY2018/19)

Time Allowed: 2 Hours

INSTRUCTIONS TO STUDENTS

1. Please write your Student Number only. Do not write your name.
2. This assessment paper contains **TWENTY** Multiple Choice Questions and FOUR Structured Questions, and comprises **Nineteen** printed pages including the cover page. The total mark is 60.
3. Students are required to answer **ALL** questions. Students should use the **OCR Form** for Multiple Choice Questions, and write the answers for each Structured Question in the space provided below each question.
4. This is a **CLOSED BOOK** assessment. Students are allowed to bring only a single sheet of A4 help-sheet for reference.
5. Students are provided with statistical tables for reference.
6. Students are permitted to use approved non-programmable electronic calculators only.

STUDENT NO: _____

This portion is for examiner's use only

Section B	Marks	Remarks
Question 1		
Question 2		
Question 3		
Question 4		
Total		

----- page left blank ----------

Section B: Structured Questions (Total 40 marks, marks as indicated for each question)

Question 1 (10 marks)

A toy factory produces three types of wooden figurines. The time required (in minutes) for preparing, assembling and packaging a unit of each type is as shown below:

Process	Type A	Type B	Type C
Preparing	60	120	90
Assembling	40	40	60
Packaging	30	20	30
Sales Price ($)	11	16	15

The factory operates on 2 shifts, with 200 workers in each shift. Each worker works 5 days a week, and 9.5 hrs each day. The ratio of time each worker works on the processes of preparing, assembling and packing is 60:23:12. How many units of each type must be produced in a week to maximise sales ($)? (10 marks)

Question 2 (10 marks)

a) The following excerpt is adapted from an article "Here's a better way to schedule surgeries", published in Kellogg Insight, Sept 5 2018.

When patients arrive at the hospital for a surgery, they probably don't think about what goes into scheduling the operating room (OR) where the procedure will happen. Optimizing Operating Room (OR) scheduling is a top priority for hospital administrators. ORs are costly to operate—$1000 per hour is typical—but they also make money. Surgical operations and associated hospitalizations typically generate about 70% of total hospital revenues.

Hospital administrators generally use a scheduling system to determine which OR will be assigned to meet a surgeon's request for a particular surgery. This process is deceptively complex because in order to optimize OR scheduling, hospitals must take three main considerations into account.

First, the cost: a fully staffed OR costs an estimated $15-20 per minute to run, or about $1,000 or more per hour. That means the fewer ORs the hospital can keep open at any given time, the greater the savings.

Second, hospital administrators try their best to honor surgeons' requests for specific OR slots. Surgeons are very important from a hospital's point of view, so the administrators work hard to respect their preferences.

The third factor is the "noisy" nature of OR requests, which arrive at irregular, unpredictable times given that patients' health problems do not occur on a regular schedule. The challenge is that the hospital doesn't know when the OR requests will arrive. It's just a call that the surgeon makes, and the hospital needs to find the space—they can't move the patient to a different hospital or surgeon.

Another unknown is the exact amount of time an upcoming surgery will take. Surgeons have a general sense of how long a given procedure takes, but there is a very wide range—it could be half the requested time or double, and hospitals need to find a way to accommodate this.

Because so many factors in OR scheduling are unpredictable and unknown, hospitals tend to prep and staff an excess number of ORs to accommodate all requests. It's a problem of excess capacity—and a costly one.

With better utilization of OR capacity, the hospital would be able to see more patients in a shorter period of time, with lower rejections or delays for surgery requests." In the long run, these savings will lead to lower healthcare costs and thus lower insurance premiums.

Describe briefly how the hospital could use the 3 types of analytics (namely descriptive, predictive and prescriptive) to solve the OR scheduling problem depicted in the case above. (6 marks)

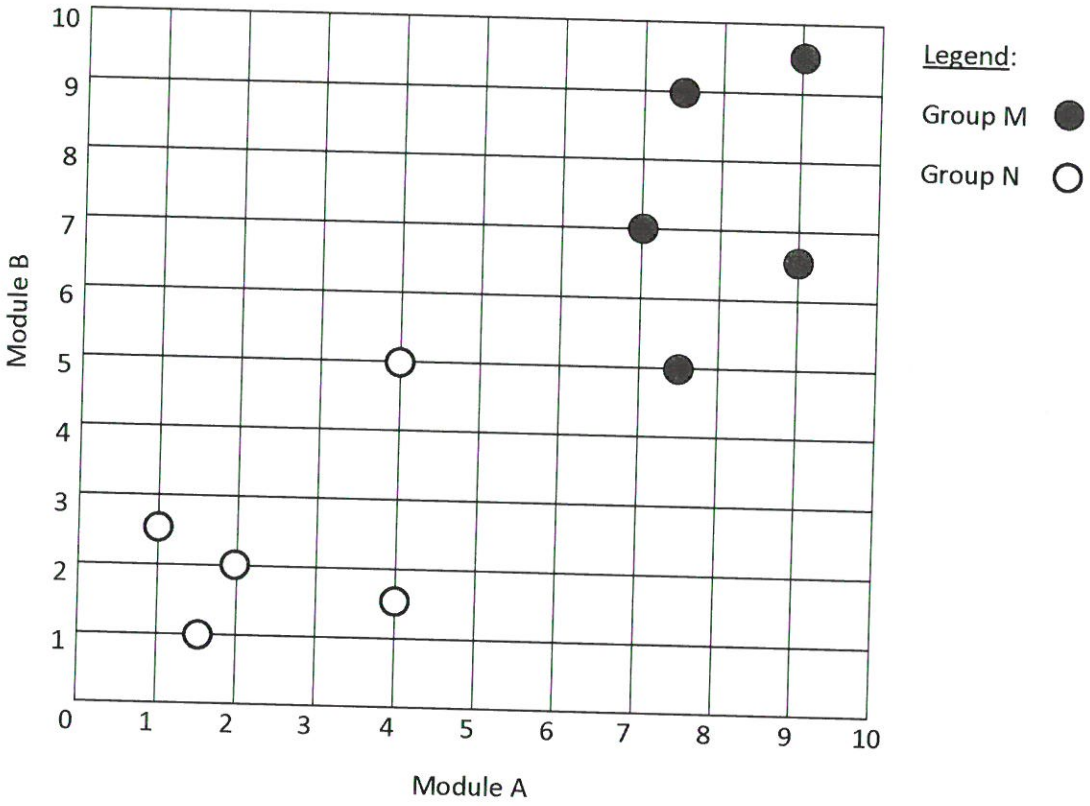
b) A paper making machine can create papers of thickness between 1.95 and 2.07mm, according to its manufacturing specifications. The manufacturing process which is normally distributed, has a standard deviation of 0.06mm.

i) If the actual mean thickness of the paper created is 1.98mm, what fraction of the papers manufactured will meet specifications? (2 marks)

ii) How small must the standard deviation be to ensure that no more than 2% of papers are non-conforming, assuming the mean is 2.01mm? (2 marks)

Question 3 (10 marks)

The following shows the mid-term test scores of Module A and Module B, by two groups of students from two different undergraduate programmes.



a) Compute the single and complete linkage clustering values of the two groups. (2 marks)

b) Estimate a value of k, and use the k-NN classifier with the Euclidean distance metric to classify the following test records: (4 marks)

Student Index	Module A	Module B	Group
Student W	6.0	3.5	
Student X	3.5	8.0	
Student Y	5.0	4.0	
Student Z	7.5	3.0	

c) Using the same estimated value of k in (b), use the k-NN classifier with the Manhattan distance metric to classify the same test records in (b). (2 marks)

d) In computing distance metric, when would Manhattan be better than Euclidean? (2 marks)

Question 4 (10 marks) *BT1101*

A study is being conducted to assess customers' perception of the customer service, value for dollar and signal strength received from their mobile carrier. The figures below present the first 13 rows of the data "Cell Phone Survey" as well as the output for a series of R commands that were ran. Use what you can gather from these figures to answer the questions that follow below.

	Gender	Carrier	Type	Usage	Signal strength	Value for the Dollar	Customer Service
1	M	AT&T	Smart	High	5	4	4
2	M	AT&T	Smart	High	5	4	2
3	M	AT&T	Smart	Average	4	4	4
4	M	AT&T	Smart	Very high	2	3	3
5	M	AT&T	Smart	Very high	5	5	2
6	M	AT&T	Smart	Very high	4	3	5
7	M	AT&T	Smart	Very high	3	4	4
8	F	AT&T	Smart	Very high	3	2	3
9	F	AT&T	Smart	Very high	4	3	4
10	M	AT&T	Smart	Very high	3	3	1
11	M	Other	Smart	Average	1	2	4
12	M	Sprint	Smart	Very high	3	5	4
13	M	Sprint	Smart	Very high	3	5	3

Showing 1 to 13 of 52 entries

```
> cps<-Cell_Phone_Survey
> ss<-cps$`Signal strength`
> value <- cps$`Value for the Dollar`
> cs<-cps$`Customer Service`

> summary(ss)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.308   4.000   5.000
> summary(value)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.423   4.000   5.000
> summary(cs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.000   3.231   4.000   5.000
> library("psych", lib.loc="~/Library/R/3.4/library")

> describe(ss)
   vars  n mean sd median trimmed  mad min max range  skew kurtosis   se
X1    1 52 3.31  1      3    3.31 1.48   1   5     4 -0.17    -0.45 0.14
> describe(value)
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 52 3.42 0.96      3    3.43 1.48   1   5     4 0.02    -0.46 0.13
> describe(cs)
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 52 3.23 0.96      3    3.26 1.48   1   5     4 -0.33    -0.02 0.13
```

```
> table(ss)
ss
 1  2  3  4  5
 2  8 20 16  6
> table(value)
value
 1  2  3  4  5
 1  6 23 14  8
> table(cs)
cs
 1  2  3  4  5
 3  6 23 16  4

> describeBy(ss,group=cps$Gender)

 Descriptive statistics by group
group: F
   vars  n mean  sd median trimmed  mad min max range  skew kurtosis   se
X1    1 18 3.06 0.8      3    3.06 1.48   2   4     2 -0.09    -1.53 0.19
------------------------------------------------------------
group: M
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 34 3.44 1.08      3     3.5 1.48   1   5     4 -0.35    -0.37 0.18

> describeBy(ss,group=cps$Carrier)

 Descriptive statistics by group
group: AT&T
   vars  n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 26 3.46 1.03      3    3.45 1.48   2   5     3  0.1    -1.22 0.2
------------------------------------------------------------
group: Other
   vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 9 2.67 1.22      3    2.67 1.48   1   4     3 -0.16    -1.75 0.41
------------------------------------------------------------
group: Sprint
   vars n mean   sd median trimmed mad min max range  skew kurtosis  se
X1    1 5  2.8 0.45      3     2.8   0   2   3     1 -1.07    -0.92 0.2
------------------------------------------------------------
group: T-mobile
   vars n mean sd median trimmed mad min max range skew kurtosis se
X1    1 2    3  0      3       3   0   3   3     0  NaN      NaN  0
------------------------------------------------------------
group: Verizon
   vars  n mean   sd median trimmed mad min max range skew kurtosis  se
X1    1 10  3.8 0.63      4    3.75   0   3   5     2 0.09    -0.93 0.2

> describeBy(ss,group=cps$Type)

 Descriptive statistics by group
group: Basic
   vars  n mean  sd median trimmed mad min max range  skew kurtosis   se
X1    1 12    3 0.85      3     3.1   0   1   4     3 -0.81     0.15 0.25
------------------------------------------------------------
group: Camera
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 19 3.26 0.99      3    3.24 1.48   2   5     3 0.15    -1.22 0.23
------------------------------------------------------------
group: Smart
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 21 3.52 1.08      4    3.59 1.48   1   5     4 -0.4    -0.49 0.24
```

```
> describeBy(ss,group=cps$Usage)

 Descriptive statistics by group
group: Average
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 23 3.13 1.01      3    3.11 1.48   1   5     4    0    -0.67 0.21
------------------------------------------------------------------------
group: High
   vars n mean sd median trimmed mad min max range skew kurtosis se
X1    1 2    5  0      5       5   0   5   5     0  NaN      NaN  0
------------------------------------------------------------------------
group: Low
   vars n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 5  3.4 0.55      3     3.4   0   3   4     1 0.29    -2.25 0.24
------------------------------------------------------------------------
group: Very high
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 22 3.32 0.99      3    3.33 1.48   1   5     4 -0.35    -0.41 0.21

> t.test(cs~cps$Gender)

        Welch Two Sample t-test

data:  cs by cps$Gender
t = -1.6452, df = 38.625, p-value = 0.108
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9764644  0.1006474
sample estimates:
mean in group F mean in group M
       2.944444        3.382353

> t.test(value~cps$Gender)

        Welch Two Sample t-test

data:  value by cps$Gender
t = -0.18455, df = 34.153, p-value = 0.8547
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6279817  0.5234065
sample estimates:
mean in group F mean in group M
       3.388889        3.441176


> aov.cs.gen<-aov(cs~cps$Gender)
> summary(aov.cs.gen)
             Df Sum Sq Mean Sq F value Pr(>F)
cps$Gender    1   2.26  2.2569   2.509  0.119
Residuals    50  44.97  0.8995
> aov.cs<-aov(cs~cps$Type)
> summary(aov.cs)
             Df Sum Sq Mean Sq F value Pr(>F)
cps$Type      2   0.49  0.2445   0.256  0.775
Residuals    49  46.74  0.9539
> aov.value<-aov(value~cps$Gender)
> summary(aov.value)
             Df Sum Sq Mean Sq F value Pr(>F)
cps$Gender    1   0.03  0.0322   0.034  0.853
Residuals    50  46.66  0.9332
> aov.value.type<-aov(value~cps$Type)
> summary(aov.value.type)
             Df Sum Sq Mean Sq F value Pr(>F)
cps$Type      2   5.26  2.6306   3.111 0.0535 .
Residuals    49  41.43  0.8455
```

a) In the table below, state the data type, the possible values and skewness (type and degree of skewness) for each of the 7 variables in the Cellphone Survey dataset. Put "NA" in the cells if the column does not apply. (2 marks)

Variables	**Data Type (e.g. categorical, ordinal, interval, ratio)**	**Possible values (state all possible values for a categorical variable or the lower and upper limit for numerical variables)**	**Skewness**
Gender			
Carrier			
Type			
Usage			
Signal Strength			
Value for Dollar			
Customer Service			

b) State and test the hypothesis for the following:

i) average customer service is the same for males and females. (2 marks)

ii) average value for dollar is the same across different "type" of phones. (2 marks)

c) Based on the survey data collected it was concluded that at least 60% of customers are satisfied (rating of 4 or 5) with the signal strength they receive. Do you agree? (1 mark)

d) Explain the difference between a prediction interval and a confidence interval. What is the 95% prediction interval for the signal strength for women respondents. Briefly describe what this 95% prediction interval tells us? (3 marks)

----- End of Paper -----