Business Data Analytics MTAT.03.319 **Name:** _____
Exam 2
18.01.2018

This exam contains 6 pages (including this cover page) and 15 questions. Total of points is 30. It is not allowed to use any materials during the exam nor consulting your peers. The duration of the exam is 2 hours. You will have 5 minutes to inspect your materials during the exam.

Grade Table (for lecturer use only)

Question	Points	Score
1	1	
2	2	
3	3	
4	1	
5	1	
6	2	
7	1	
8	4	
9	3	
10	2	
11	3	
12	1	
13	3	
14	1	
15	2	
Total:	30	

1. (1 point) Mark box if true.
   A. Descriptive BA is about "What happened" question
   B. Descriptive BA is about "What will happen" question
   C. Inclusive BA is about "What to include" question
   D. Prescriptive is about "How to improve" question

E. Exclusive is about "How to improve" question

F. Predictive is about "What will happen" question

2. (2 points) The Megamegacorp is (fictitious) big corporation selling appliances. Among all the customers who bought their products online they asked to fill the following survey:

   1. Level of education: Pre-school, Basic, Secondary, Higher
   2. Do you have a loyalty card: Yes, No
   3. Choose your favorite brand: 1) KnivesCo 2) FluffyFluff, 3) FilipBosch 4) Other
   4. What was the amount of money spent on your last order? ....

   Please, specify for each of the questions, what data type is the answer to it, where:

   A. binary

   B. nominal

   C. ordinal

   D. discrete

   E. continuous

   Your answer format should be something like this: 1)A, 2)B, 3)C, 4)A

3. (3 points) As a senior analyst of the Megamegacorp you get the following report from one of the interns:
   In December 2017 805 customers bought refrigerators. 75% bought it online. The average amount spent of online customers is 5.76 times higher (mean of offline customers is 219.4 while online customers is 1263.5).
   You found it somehow suspicious and decided to double check the answers. The results were the following:

```
> summary(order_total_offline)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  101.1   149.1   189.6   219.4   267.9   685.7
> summary(order_total_online)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   102.9    145.4    179.8   1263.5    267.2 634000.0
```

   Were your suspicions right? Can you say that online customers are 5.76 times more profitable? Choose the closest answer to your investigation.

   A. No, the Minimum amount indicates that the customers are equally profitable.

   B. All is correct. The intern did a good job!

   C. No, the 5.76 times is not correctly calculated, it should be 8 times.

D. Online customers 3rd quartile and Max are very different. It must be due to the outliers. Therefore, mean is not a correct metric to use to measure profitability.

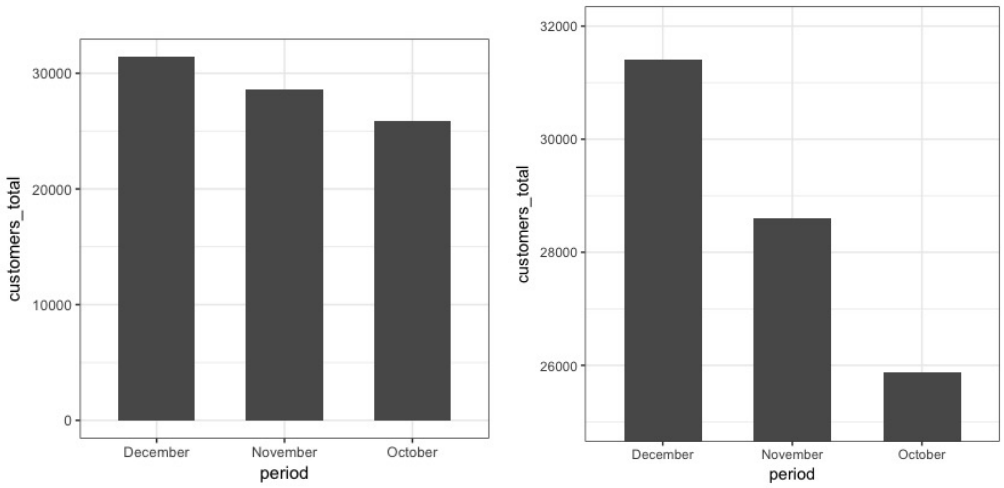
E. Online customers 1st quartile and Max are very different. It must be due to the outliers. Therefore, mean is not a correct metric to use to measure profitability.

F. No, it is not correct as means are different from ones that the intern stated.

G. Online customers 3rd quartile and Max are very different. Online customers are profitable.

H. As the size of offline and online groups are different, we can't use mean.

4. (1 point) The intern also provided you with two figures and asked your help. Which of the plots you'd choose and why?



(a) (b)

A. plot (a) as it does not skew the data showing the adequate difference

B. plot (b) as it highlights the difference between months

C. plot (a) as the last y-axis tick is closer to the maximum

D. Both plots are bad

5. (1 point) RFM analysis ranks customers by considering ..... of their orders.

6. (2 points) One of the customers of the Megamegacorp is a frequent buyer. This client just made his regular purchase. However, he usually spends little money and none of his purchases was expensive. Under the given circumstances, what is his most likely RFM score:

A. 233

B. 115

C. 511

D. 555

7. (1 point) Supervised learning differs from unsupervised learning in that supervised learning requires

   A. at least one input feature.

   B. input features to be categorical.

   C. at least one output feature.

   D. output features to be categorical.

8. (4 points) You have a dataset with two dimensions, customer loyalty score and customer value score, and the following dataset for 3 customers:

```
> dt
  customer_id loyalty_score value_score
1         178             2           4
2         163             5           7
4         112             1           1
```

   Perform k-means. Use $k = 2$ and initialize seeds of (3,4) and (5,2). Use Euclidean distance. Write your calculations and assign each customer to one of each clusters, as well as state how many iterations you did before converging to the result:

9. (3 points) Consider the following confusion matrix:

		Actual class: Positive	Actual class: Negative	Total
Predicted class	Positive	50	30	80
	Negative	25	15	40
	Total	75	45	120

   Calculate Precision and Recall based on these numbers.

10. (2 points) The Megamegacorp sales team wants to test a claim that if they change a page stating "Discounts!" to "Discounts! Only 9 hours left" then the average number of clicks per day will **increase**. Which hypothesis you formulate to test this claim?

    A. $H_0 : \mu_A = \mu_B$ vs. $H_0 : \mu_A < \mu_B$

    B. $H_0 : \mu_A = \mu_B$ vs. $H_0 : \mu_A \neq \mu_B$

    C. $H_0 : \mu_A \neq \mu_B$ vs. $H_0 : \mu_A = \mu_B$

    D. $H_0 : \mu_A > \mu_B$ vs. $H_0 : \mu_A = \mu_B$

11. (3 points) In the Megamegacorp you also have a special feature – an auction, where customers sell their appliances and bid how much they are willing to pay. Your team built a regression model for the Megamegacorp predicting the monetary value (y) of the appliance. It is based on 1) initial price, 2) age of the product, 3) number of posted pictures of the product, 4) seller response time (in minutes). For each type of appliance you have a separate model. Let's say that for the vacuum cleaners your model looks like this (all features significant, index indicates the feature from the list):

$$\hat{y} = 40 + 3 \times x_1 - 2 \times x_2 + 1.18 \times x_3 - 1.8 \times x_4$$

    Mark all that apply.

    A. by increasing initial price by 1 euro, the final price increases in average by 40

    B. by increasing number of pictures by 1, the final price increases in average by 1.18

    C. number of pictures and initial price increase the estimated value of the product, while age of the product decreases the estimate.

    D. an increase in response time by 1 min, decreases the estimated value in average by 1.8 euros.

12. (1 point) Calculate the estimated price of a vacuum cleaner if the initial price of a vacuum cleaner is 100, the product is 4 years old, number of pictures is 10, and the seller response time in average is 30 mins.

13. (3 points) The Metametacorp has recently acquired three different small companies that run popular websites for appliances' reviews. Users of these websites compared different brands of same products. Now, you want to combine information about the reviews in order to build collaborative filtering. However, the problem is that the first website uses ranking with stars from 1 to 5, the second – the score from 1 to 10, and the third one gives the score as a percentage from 1 to 100. Which of the statements is true:

    A. You can combine all three datasets without any changes.

    B. You can combine all three datasets and then (**after merge**) perform mean normalization and scaling. All scores will be on the same scale.

    C. You **first** scale each of the datasets to be on the scale from 0 to 1 and then merge them together. ll scores will be on the same scale.

    D. It is not possible to combine these datasets.

14. (1 point) In order to be used as input for process mining, an event log must contain at least which of the following fields? (Mark all that apply)

    A. case id

    B. timestamp

    C. status

    D. activity

    E. event id

    F. resource

15. (2 points) Imagine you have an event log of a loan application process. Every case represented in the log corresponds to one loan application. Some loan applications lead to loan offers (i.e. the customer received a loan offer). Others lead to a rejection (i.e. the customer did not receive a loan offer). If a customer receives a loan offer, the customer may accept the loan offer or reject it. If the offer is accepted, the loan is paid out (disbursed) by the bank. All the events related to the acceptance or rejection of loan applications, the acceptance or rejection of the loan offers, and the loan disbursement (payment to the customer), are recorded in the event log. Imagine you want to know if the cases where the customer accepts the offer take more time on average than those where the customer does not accept the offer. Which of the following filters would you use?

    A. event filter (i.e. an attribute filter over the "activity" field)

    B. endpoint filter

    C. follower filter

    D. performance filter

    E. none of the above