|  |  |
| --- | --- |
| Part 1 (36) |  |
| Part 2 (8) |  |
| Part 3 (56) |  |
| Total (100) |  |
| Do not mark - for grading purposes only! | |

**ITEC 620**

AMERICAN UNIVERSITY

MIDTERM EXAM (**PRACTICE**)

INSTRUCTIONS

There are three parts of the exam, totaling 100 points:

9 Multiple Choice questions (4 points each)

1 Short Essay question (8 points)

4 Problems (8, 20, 8, and 20 points)

You may use one double-sided sheet of notes only. All electronic devices are prohibited, except for a basic calculator (scientific or graphing calculator is fine).

Exams are randomized in several different ways. You will not have exactly the same exam as the people around you.

Good luck!

***Please note that this practice exam is NOT intended to be a carbon copy of the actual midterm. Its purpose is to give you a sense of the length, scope, and format to expect.***

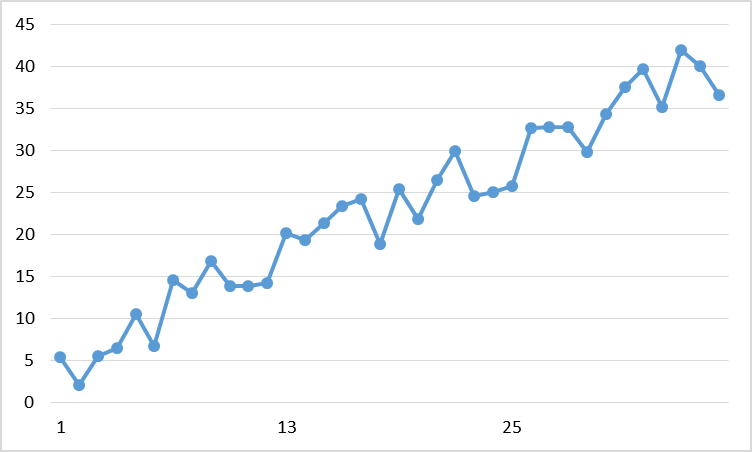
**Part 1: Multiple Choice Questions**

1. Which of the following is NOT part of the analytics cycle?

A) Formulate the question/problem  
B) Identify and collect the available data related to the question  
C) Determine what outcome you hope to find  
D) Write up conclusions and recommendations

2. Which of the following is an example of descriptive analytics?

A) Determining the correlation between daily website traffic and number of server outages  
B) Finding the best possible server configuration  
C) Forecasting the number of website visitors between 12-4 AM on Thursday  
D) Predicting which banner ad a user will be the most likely to click

3. Which of the following would be the most appropriate forecasting method for the time series shown in the chart below? (The time series contains monthly data.)  
  


A) Single exponential smoothing  
B) Double exponential smoothing  
C) Holt-Winters  
D) Moving average

4. Which of the following is generally true about large data sets (say, with *n*=8000)?

A) They cannot be analyzed without cloud computing and professional versions of software.  
B) Because the sample size is large, we can assume that no biases exist in the data.  
C) Traditional descriptive statistics can detect all relevant patterns and relationships in the data.  
D) Some analytics methods are more useful with these datasets than they would be with less data.

5. X and Y have a correlation of 0.8. Which of the following is true?

A) When X is higher than average, Y tends to be higher than average  
B) When X is higher than average, Y tends to be lower than average  
C) X and Y are not related in any substantial way  
D) Changing X will cause Y to change as well

6. Which of the following is true of outliers?

A) Every data set has at least one outlier  
B) An outlier is usually a data error  
C) We can remove outliers without affecting the result of analytics methods  
D) Outliers affect the mean of a variable more than the median

7. What does the command: “Salaries[8]” do in R? (You can assume the Salaries object contains a vector of salaries.)

A) Display the first eight salaries  
B) Display the eighth salary  
C) Delete all salaries after the eighth one  
D) Set the value of Salaries to eight

8. Which of the following is true of lift ratios in association rules?

A) A lift ratio close to zero indicates no noticeable relationship between the premise and the conclusion  
B) A lift ratio of 0.8 means that, given the premise, the probability of the conclusion is 0.8  
C) A lift ratio of 3 means that the premise makes the conclusion 3 times more likely  
D) A lift ratio of 1 means that the conclusion is true for all of the data points

9. Which of the following is NOT a benefit of partitioning a time series to compare forecasting models?

A) We can compute a measure of predictive accuracy for each model  
B) It makes it less likely that we will end up making predictions with an overfit model  
C) It increases the number of data points that we can use to build each model  
D) We can evaluate the models based on data points that were not used to create them

**Part 2: Short Essay Questions**

1. (8 pts.) Choose any real-world organization. Describe clearly and specifically two ways in which this organization could use descriptive analytics methods to perform more efficiently or effectively.

**Part 3: Problems**

**1**. (8 pts.) We have observed the following time series (in chronological order):

18, 26, 26, 16, 30, 27, 30

a) (4 pts.) Using a 3-period moving average, what is the prediction for the next value in the time series?

b) (4 pts.) Using a single exponential smoothing model with alpha=0.5, the prediction for the most recent period was 26. What is the prediction for next period?

**2**. (19 pts.) The following is a portion of association rules output in R for transactions at a local coffee shop:

lhs rhs support confidence lift

[1] {Sandwich} => {Drip Coffee} 0.02 0.20 0.80

[2] {Muffin} => {Drip Coffee} 0.08 0.40 1.60

[3] {Latte} => {Sandwich} 0.01 0.05 0.50

[4] {Drip Coffee} => {Muffin} 0.08 0.32 1.60

[5] {Muffin} => {Latte} 0.06 0.30 1.50

a) (4 pts.) If a customer bought a sandwich, what is the probability that that customer also bought a drip coffee?

b) (4 pts.) Is the overall proportion of customers who bought drip coffee higher or lower than the answer to part a?

c) (4 pts.) What is the support of “Latte”? That is, what proportion of overall customers bought a latte? (Hint: use the last rule in the table)

d) (4 pts.) While most items can be made quickly, lattes and sandwiches each require substantial individual attention from an employee. Why is the low lift ratio of the third association rule in the table helpful, from the perspective of a manager running the coffee shop? (Please limit your answer to 40 words or less.)

e) (4 pts.) If there are 1000 transactions in the data set, how many of these transactions include both a muffin and a latte?

**3**. (8 pts.) The following output shows a centroid table produced by running k-means clustering in R. Each data point is a company for which we have three variables: the total number of employees at the company, the valuation (estimated total $ value) of the company, and the number of countries in which the company has at least one office.

Number of Employees Valuation ($B) Global Presence (# of countries)

1 3300 0.61 8.32

2 51700 50.37 137.85

3 46500 48.24 16.94

4 17200 16.22 46.61

a) (4 pts.) Briefly (1-2 sentences), describe the type of company that would be in Cluster 2.

b) (4 pts.) American Conglomerates is one of the largest companies in the data set, but has no offices outside North America. Which cluster is it in?

**4**. (20 pts.) Each of the following blocks of R code is not functioning as intended due to one line with a mistake. For each one, identify the line of code with a mistake, and rewrite it correctly.

a) (5 pts.) Running k-means clustering with *k*=4 on a dataset of athletic conferences:

Conferences.norm <- scale(Conferences)

set.seed(12345)

Conferences.kmclusters <- kmeans(Conferences, 4, nstart=10)

b) (5 pts.) Calculating the mean of the first five columns of a customer dataset:

for (i in 1:5) {

col\_mean <- mean(Customers[i,])

print(paste(“The mean of column”,i,”is”,col\_mean))

}

c) (5 pts.) Creating and sorting a set of association rules for a recipes dataset:

rules <- apriori(recipes.binary)

rules.sorted <- sort(rules, by=”lift”)

inspect(recipes.binary)

d) (5 pts.) Creating a single exponential smoothing model for a GDP time series of annual data starting in 1950:

GDP\_ts <- ts(GDP,start=1950)  
GDP\_ts\_SES <- HoltWinters(GDP\_ts)