Cairo University
Faculty of Computers and Artificial Intelligence

Midterm Exam

Department: Operations Research and Decision Support
Course Name: Data Analytics
Course Code: DS342
Instructor(s): Dr.Marwa Sabry

Date: November 29, 2021
Duration: 1 hour
Total Marks: 40 Marks

تعليمات هامة

- حيازة التليفون المحمول مفتوحا داخل لجنة الإمتحان يعتبر حالة غش تستوجب العقاب وإذا كان ضرورى الدخول بالمحمول فيوضع مغلق فى الحقائب.
- لا يسمح بدخول سماعة الأذن أو البلوتوث.
- لايسمح بدخول أي كتب أو ملازم أو أوراق داخل اللجنة والمخالفة تعتبر حالة غش.

- **Save your file under "ID–Your Name".**
- **Save your work continuously on the Specified Folder. It is YOUR RESPONSIBILITY to make sure that your work is saved.**
- **Attempt all problems.**
- **At the end of the exam, ask the TA responsible in your lab to check that the file is saved and available.**

Question 1 ***[10 marks]***

State which of the following statements is True(T) / False(F)

F 1. Phone numbers, Social Security numbers, and zip codes are typically treated as numerical variables.

T 2. Problems in data analysis where we want to compare a numerical variable across two or more subpopulations are called comparison problems.

F 3. Correlation and covariance can be used to examine relationships between numerical variables and categorical variables that have been coded numerically.

T 4. Although it is relatively easy to collect data, it can be more challenging to understand what the data mean.

F 5. If the standard deviation of X is 15, the covariance of X and Y is 94.5, the coefficient of correlation r = 0.90, then the variance of Y is 7.0.

T 6. A few ways to enhance the readability of a spreadsheet model is to use a clear, logical layout of the overall model, and to use clear headings for different sections of the model and for all inputs, decision variables, and outputs.

F 7. A distribution with a high kurtosis has almost all of its observations within three standard deviations of the mean.

T 8. If a histogram of a data set is symmetric and bell shaped, with a mean of 75 and standard deviation of 10. Then, approximately 95% of the data values will be between 55 and 95.

F 9. Power Pivot is needed to create a Data Model.

F 10. The feature that sets Power Pivot apart from regular Excel pivot tables is the relatively new language, Structured Query Language (SQL) for performing calculations.

Question 2 ***[15 marks]***

Choose the correct answer from A, B, C, or D

1. Which of the following would *not* be included under data analysis?
   a. Measuring uncertainty
   b. Data description
   c. Data inference
   d. Search for relationships
2. Tables used to display counts of a categorical variable are called
   a. crosstabs.
   b. contingency tables.
   c. either crosstabs or contingency tables.
   d. neither crosstabs nor contingency tables.
3. Examples of comparison problems include
   a. salary broken down by male and female subpopulations.
   b. cost of living broken down by region of a country.
   c. recovery rate for a disease broken down by patients who have taken a drug and patients who have taken a placebo.
   d. all of these choices.
4. Scatterplots are also referred to as
   a. crosstabs.
   b. contingency charts.
   c. X-Y charts.
   d. all of these choices
5. Correlation and covariance measure the
   a. strength of a linear relationship between two numerical variables.
   b. direction of a linear relationship between two numerical variables.
   c. strength and direction of a linear relationship between two numerical variables.
   d. strength and direction of a linear relationship between two categorical variables.
6. Coding males as 1 and females as 0 in a data set illustrates the use of _____ variables.
   a. nominal
   b. dummy
   c. numerical
   d. Ordinal
7. One characteristic of "paired variables" is that
   a. one variable is a negative value and the other is a positive value.
   b. both variables are positive values.
   c. each variable has the same number of observations.
   d. each variable has a different number of observations.
8. Changing the location of fields in a pivot table is known as
   a. slicing.
   b. dicing.
   c. sorting.
   d. pivoting.

9. Within the Power Pivot window, you can choose to hide individual columns or even entire tables. Why might you want to do this?
   a. It prevents users from adding fields to pivot tables that would certainly lead to errors.
   b. It creates less confusion for users, so that they won't be tempted to use fields in pivot tables that wouldn't lead to interesting results.
   c. It ensures that fields unrelated to other fields won't be used in pivot tables.
   d. Its only use is to hide primary and foreign key fields so that they won't be used in pivot tables.
10. Suppose you try to create a pivot table from multiple tables stored in a Data Model and, when you check a particular field to be placed in the pivot table, you get a warning about a missing relationship. Which of the following does this *not* imply?
   a. You are trying to use fields from unrelatable tables, so no matter what you do, your pivot table results will be wrong.
   b. You can open the Power Pivot window, create the missing relationships, and try building the pivot table again.
   c. You can click the Auto-Detect button in the PivotTable Fields pane, and the chances are that Excel will correctly create the missing relationship.
   d. You can click the CREATE button in the PivotTable Fields pane, which allows you to manually create the missing relationship.
11. Which of the following is true?
   a. When entering an expression of text into an excel function, it must be enclosed in double quotes.
   b. A spreadsheet model should always include input numbers, rather than cell references, in formulas.
   c. If we enter A1:A5 as part of an Excel function, this refers to cells A2, A3, and A4…the cells that are between A1 and A5, exclusive.
   d. All of these statements are true.
12. Which of the following is *not* a primary advantage of Microsoft's self-service BI tools?
   a. They allow ordinary Excel users to perform powerful data analysis without needing help from a corporate IT department.
   b. The tools let you analyze the data in Excel. However, you must first create a relational database in a database package such as Access.
   c. To analyze large relatable tables in Excel, the usual practice of employing multiple VLOOKUP functions is no longer required.
   d. These tools allow you to analyze data sets in Excel that go well beyond Excel's approximately one million row limit.
13. In Excel, the model outputs are
   a. the numeric values that result from combinations of inputs and decision variables through the use of logical formulas.
   b. useful for making formulas more readable.
   c. the variables a decision maker has control over to obtain the best solutions.
   d. useful for finding a particular value based on a comparison.

14. Where will you find "time" on a time series graph?
    a. horizontal axis
    b. first column
    c. vertical axis
    d. last column
15. Categorizing a numeric age variable as "young," "middle-aged," and "elderly" is an example of
    a. counting.
    b. ordering.
    c. quantifying.
    d. binning.

Question 3 *[15 marks]*

Referring to the data file located on your desktop named "Shirt Orders.mdb", it is an Access database file with three related tables: **Customers, Orders, and Products. You are requested to perform the following:**

a. Find all records from the Orders table where the order was placed in 2015 or 2016, the product ID is from 1 to 5, the customer ID is not 7, and the number of units ordered is at least 75. Import all fields in the Orders table for each of these records into an Excel table (not a Data Model). *[5 marks]*

b. Create a table in Excel that contains the order date, customer name, and product description for all orders that satisfy conditions on orders, products, and customers as follows:
    1. Use Power Query to import all three tables into Excel as tables (not a Data Model). Then use Query Editor is to remove the following columns: Street, City, State, Zip, Phone for **customers**, Discount for **orders**, and UnitPrice for **products**. Save the imported data. *[2 marks]*
    2. Add three new fields, Customer Name, Product Description, and Gender, to the **Orders** table in Excel and use VLOOKUP functions to fill them. *[3 marks]*
    3. Filter the Orders table as follows: units ordered ">75", "Both" for gender, and both "Threads" and "Shirts R Us" for customer. *[2 marks]*
    4. The company would like to know how many units of the products designed for each gender category (men, women, and both genders) were sold to each customer during each quarter of the past five years (Q1 of 2012 through Q4 of 2016). Use Power Query to import the appropriate data into an Excel pivot table report (not a Data Model or a table) and then manipulate the pivot table to show the required information. *[3 marks]*

Good Luck

Dr. Mo