

### Assignment Overview

Students will complete an **individual assignment** in data management and summarization using Microsoft Excel. You will be given access to an unorganized dataset and will need to organize and summarize the data in Excel using various formulas, functions, tables, and charts.

**Due date:** November 22<sup>th</sup>, 2019 by 10:00 AM on D2L

**Format of submission:** Excel spreadsheet (.xls or .xlsx file)

**Evaluation:** This assignment is worth 20% of your final grade

### Dataset

Health care institutions are required under the Ontario *Health Protection and Promotion Act* (HPPA) to monitor staff and patients for symptoms of gastroenteric and respiratory infections (City of Toronto, 2019). Institutions are required to detect, and report suspected and confirmed infections to their public health unit. An outbreak of such infections is an increase in the rate of infections above what is expected. Using open-source data from the City of Toronto (please see below link), you will be required to download and analyze institutional outbreak data following the instructions provided below.

*Open Data Catalogue: Outbreaks in Toronto Healthcare Institutions (Outbreaks in Toronto Healthcare Institutions, 2017 AND Outbreaks in Toronto Healthcare Institutions, 2018)*

<https://open.toronto.ca/dataset/outbreaks-in-toronto-healthcare-institutions/>

### Assignment Instructions (20 Marks Total)

#### Part A (10 Marks)

1. Download and merge both “outbreaks-in-toronto-healthcare-institutions-2017” and “outbreaks-in-toronto-healthcare-institutions-2018-year-to-date” data as an *xls* or *xlsx* spreadsheet file. **(1 Mark)**
  - a. Format and merge both files into one Excel workbook with 2 sheets for the individual years (2017 and 2018) and one sheet for the merged data. Label the sheets accordingly.
  - b. Label the file starting with your last name, followed by your student number, and then the course code. Separate these with an underscore (e.g. NAME\_0123456789\_POH103).
2. By default, the data are sorted by “Institution Name”. Resort your merged data worksheet alphabetically by “Type of Outbreak” then by “Outbreak Setting”. **(1 mark)**

3. For this assignment we will only consider “Causative Agent-1” and not “Causative Agent-2”. Create a new column to the right of “Causative Agent-1” called “Primary Causative Agent”. Copy the contents of “Causative Agent-1” into this column, and replace all Influenza A entries to read as “Influenza A” without the subtypes. Similarly, replace all Parainfluenza entries to read as “Parainfluenza” without the subtypes, and replace all *Clostridium difficile* entries to read as “Clostridium difficile” without the toxin types. Note: leave the “Influenza A and B” and “Influenza A, Parainfluenza” combined entries as is. **(1 Mark)**
4. Insert a column after the “Date Declared Over” column called “Outbreak Duration”. For each outbreak in the spreadsheet, calculate the duration of the outbreak in this column as the difference between the outbreak start and end dates. Ensure to convert the formatting of the outbreak duration column from date to number. **(1 Mark)**
5. Create a new column to the right of the new “Outbreak Duration” column called “Outbreak Risk”. In this column, use the IF function to categorize the outbreak duration as high or low risk. For durations over 30 days, categorize them as “high risk”, otherwise as “low risk”. **(2 Marks)**
6. On a new worksheet titled “Tables”, create a table calculating the Maximum, Mean, Median, Minimum and Mode of the outbreak duration values. **(2 Marks)**
7. In a new row in the “Tables” worksheet, insert a function to count the number of “high risk” outbreaks from your “Outbreak Risk” column. In another row, create a function to calculate the average outbreak duration length for enteric outbreaks. **(2 Marks)**

#### **Part B: Pivot Tables and Graphs (8 Marks)**

8. Create a pivot table to analyze the number of outbreaks by month and year of onset to identify which months and years had higher rates of outbreaks. **(3 Marks)**
  - a. Create a pivot table in a new worksheet called “Pivot Table 1” to calculate the number of outbreaks by year and month. Use the “Date Outbreak Began” field to generate the month and year information in the Pivot Table. Note that if using a newer version of Excel, when you add this field to your Pivot Table it may automatically generate a “Year” and “Quarters” field. If so, you can remove the “Quarters” field. If these fields are not automatically generated, right click on any date in the Pivot Table and click “Group”, then highlight and select Month and Year. Months are shown automatically when you expand the years in the Pivot Table. You can use any field to generate the count of number of outbreaks in the dataset.
  - b. Create a vertical clustered column chart (graph) using the data from the Pivot Table to visually display the data. Ensure to label the chart and its x-axis and y-axis appropriately.

- c. In each year, which month had the highest number of outbreaks? Write a sentence in a cell beside or below your Pivot Table to indicate this answer in sentence form.
9. Create another Pivot Table in a new worksheet called “Pivot Table 2” to analyze the number of outbreaks by type of outbreak and the outbreak setting to identify which setting had higher rates of different types of outbreak. **(3 Marks)**
  - a. Create a Pivot Table to calculate the frequencies of outbreaks by type and setting. Set up your table so the type of outbreak is shown as columns and outbreak setting as rows.
  - b. Create a horizontal stacked bar chart (graph) to visually display the data from this Pivot Table.
  - c. Which outbreak setting reported the highest number of outbreaks of each type and overall? Write your answer in sentence form in one of the cells beside or below the Pivot Table.
10. Create one more Pivot Table in a new worksheet called “Pivot Table 3” to analyze the number of outbreaks and their average duration by the primary causative agent. **(2 Marks)**
  - a. Create a Pivot Table to calculate both the number of outbreaks and the average duration of outbreaks by the primary causative agent.
  - b. Which agent caused the most outbreaks? How many? Which agent caused the longest average outbreak duration and what was the duration value? Write your answers in sentence form in one of the cells beside or below the Pivot Table.

**Formatting (2 Marks)**

- ✓ Ensure to leave the formulas or functions used as visible
- ✓ Remember to rename the worksheets accordingly
- ✓ Graphs must be labelled with titles and axis labels
- ✓ Please show your work