Exercise 2 Excel Hints

1. Background
   1. Begin by saving the CSV file provided as an Excel workbook.
2. Data Cleaning and Summary Statistics

**Tip:** Create a new tab for your summary statistics. Converting your raw data into an Excel table (Insert → Table) can make formulas and analysis easier. Make sure your table has headers!

* 1. Heart Rate
     1. Setting up the summary table: Use formulas to reference the columns and manually add the statistics (mean, median, etc.) alongside the table.

A screenshot of a computer

AI-generated content may be incorrect.

* + 1. Convert text data to numeric:
       - If Excel treats numbers as text (e.g., sorting doesn’t work), first replace nan values with blank spaces using Find & Replace.
       - Then, highlight the column and use Data → Text to Columns to convert to numeric values.
    2. Calculating Statistics:
       - Use formulas like =STAT(Table1[COLUMN]) or in the case of max it would be =MAX(Table1[Heart Rate (BPM)]). Try identifying the other function names by typing in what you might expect them to be called.
       - For standard deviation, choose the standard deviation of the sample (STDEV.S).
    3. **Important:** Before you start, make sure you create a new tab for your cleaned data table. DO NOT MAKE CHANGES TO THE ORIGINAL DATA TABLE! If you do, then your summary statistics will change as you update the table.
    4. Cleaning the data:
       - Identify outliers by sorting or filtering the column along with your calculated summary statistics.
       - Options for handling outliers:
         1. Use Find and Replace or Sorting to manually remove or impute the value.
         2. Formulas. Note that if you’re going to use a formula, make sure the tables are sorted in the same way (I would recommend sorting based on the User ID column).

Replace with N/A:   
=IF(AND(A1>=##, A1<=##), A1, NA()) where A1 is the cell in the original data table and the two numbers would be the bounds you’d want to filter by.

Replace with Blank

=IF(AND(A1>=10, A1<=100), A1, “”)

Replace with column average (absolute reference $ required, i.e. if the average is in E4, you’ll want to do TAB\_NAME!E$4):   
=IF(AND(A1>=10, A1<=100), A1, CELL\_WITH\_AVERAGE)

* + - * 1. To fill the formula down the column: CTRL/CMD + SHIFT + ↓ then CTRL/CMD + D depending on if you’re a Windows or Mac User. You might need to hit ↓ several times until you reach the end of your table.
    1. Create a second summary table for the statistics on the cleaned data.
  1. Blood Oxygen Levels
     1. Similar process to above.
     2. If you want to only filter based on a single condition, you can use IF without the AND keyword, i.e.

=IF(A1<100, A1, “”) where A1 is the cell from the column in the original table

* 1. Step Count
     1. Similar process. Remember to convert the column to numeric data first!
  2. Sleep Duration
     1. Similar process. Remember to convert the column to numeric data first!
  3. Activity Level
     1. Convert the column first.
     2. Check the drop-down filter for inconsistencies (typos, variations).
     3. Notice how there’s some spelling or typing errors, to standardize this, we could either use Find and Replace or another IF formula. Instead of using the AND keyword, we’ll instead use the OR keyword, i.e.

= IF(OR(A2="Active", A2="Actve"), "Active",   
IF(OR(A2="Inactive", A2="Inactiv"), "Inactive",   
IF(OR(A2="Pending", A2="Pendng"), "Pending", A2)))   
  
where A2 is the cell in the column in the original table and each OR statement describes the two possible values that may be present for each self-reported activity level.

* 1. Stress Level
     1. Convert the column first.
     2. Use Find & Replace or an IF formula if you want to impute “Very High” values.

1. Data Analysis
   1. Does Cleaning make a difference?
      1. Select the columns you want to visualize and go to Insert → Chart.
      2. Customize under Chart Design → Add Chart Elements (titles, labels).
      3. You may need to adjust the axis bounds. You can do this by double clicking on the axis you’d like to adjust. Under Axis Options → You can set the upper and lower bounds of the range.
      4. To adjust the size of the points, you can double click or right click on the points to access format options for the data series. To resize: Under Format Data Series → Fill & Line → Marker Options. Feel free to play around with the options.
   2. Comparing Step Count Across Activity Levels
      1. Excel’s multi-category boxplots can be a bit tricky!
         * Check out this [tutorial](https://support.microsoft.com/en-us/office/create-a-box-plot-10204530-8cdf-40fe-a711-2eb9785e510f) for one option to make one.
         * Option 2: Create a new sheet with activity levels as column headers. To filter items for each activity level, you can use the formula:   
           =FILTER(Table#[Step Count], Table#[Activity Level] =VALUE) where Value is Sedentary, Active, or Highly Active depending on the column.
      2. Histograms:
         * Reshape the data as described above so that there are three columns (one for each activity level Sedentary, Active, Highly Active). Populate each column with the relevant step count values.
         * Generate separate histograms for each group and align the x-axis ranges.
      3. **Remember you only need to do a histogram or a boxplot, not both!**
      4. Optional:
         * Compare groups quantitatively using a one-way ANOVA:
           1. [ANOVA External Tool](https://www.statskingdom.com/180Anova1way.html) - where you can copy and paste the values for each activity level.
           2. Excel’s built-in ANOVA function
   3. Relationship between Sleep Duration and Stress Level
      1. Use =CORREL(column1, column2) to calculate correlation.
      2. Note, that this function requires that there are not any N/A values in the columns. Depending on how you decide to deal with missing values in earlier steps, you may need to remove them. If so, you can search for N/A values using =NA() in the find parameter and replace it with blank values.