# Exercise 2: Communicating a Data Analysis

## Background

You’re a data analyst for *Steppr,* a company that sells smartwatch fitness trackers (i.e. watches that count steps and monitor other personal health information). You’re being tasked to investigate two questions:

1. Is there a significant difference in the number of steps taken across self-reported activity levels among participants?
2. What is the correlation between sleep duration and user reported stress levels?

Open the dataset “healthcare\_watch\_data.csv” in either Python ([using the provided notebook](https://dsci-b215-hub.brynmawr.edu/hub/user-redirect/git-pull?repo=https%3A%2F%2Fgithub.com%2Fampivirotto%2FDF_fall25&branch=main&urlpath=lab%2Ftree%2FDF_fall25%2FExercise2_DataCleaning%2Fexercise2_pythonVersion.ipynb) in the JupyterHub) or Excel.

Take a quick look at the data and describe what you observe:

* What variables are included in the columns?
* How many participants (rows) are there?
* What do you notice about the data in each column (e.g., data types, missing values, ranges, unusual patterns)?

## Data Cleaning and Summary Statistics

One of the things that you might want to communicate with your findings is the summary statistics for each variable in the dataset. This particular dataset has 7 columns: User ID, Heart Rate (BPM), Blood Oxygen Level (%), Step Count, Sleep Duration (hours), Activity Level, and Stress Level.

Before we investigate the two questions, we want to take some time to clean and summarize the data.

### 2a. Heart Rate

Let’s start with Heart Rate. When we do a quick visual inspection of this column, you might notice abnormally high values like 240 BPM. Let’s find out what the range of values for this variable is in our dataset.

First, we’re going to calculate the **minimum** and **maximum** values. (Hint: In both python and excel you can use the min() and max() functions!) While we’re doing this, let’s also calculate **mean (average)**, **median**, and **standard deviation** of the sample. Input these values in the ***Pre-cleaning Values*** table in 2g.

Do these values make sense? If not, what might explain these unusual values (e.g., device error, data entry mistake, true extreme case)?

When working with real-world data, it’s normal to find missing or unusual values. Before analyzing your dataset, it’s important to decide how to handle these so your results are accurate and fair.

**Missing values** occur when some information isn’t recorded (for example, a participant didn’t wear their smartwatch for some portion of the study). You can handle missing data in a few ways:

* Remove rows that have missing values (if only a few are affected)
* Fill in missing values with an estimate such as an average, median, or mode for that variable (this is also called imputation)
* Leave them as missing if removing or filling them might distort the data or bias the results

**Outliers** are data points that are much higher or lower than the rest of the distribution of values (i.e. a heart rate of 250 BPM when normal range is 60-100 BPM). Some outliers are real, while others may be errors. You can:

* Check whether an outlier makes sense or seems like a mistake
* Remove extreme values that are clearly unrealistic (you could then leave them as null values, impute values to replace the extreme values, or cap the values at the reasonable minimum or maximum value)
* Compare your results with and without outliers to see how much they affect the mean and median values **(Note: this is why we’re keeping track of pre-cleaning and post-cleaning summary statistics!)**

In all cases, make sure to note what choices you made and why. There’s rarely one “right” answer! The important thing is to show that you thought carefully about your data and your choices!

Okay, now it’s time to start cleaning our data! Make a copy of your data table – for both python and excel. As we make changes only update the copied table so that we have an original version of the data table.

For heart rate, we’re going to impute values for our outliers and missing data. First let’s deal with the outliers – what’s a reasonable minimum and maximum values for heart rate?

For any values above or below what you’ve deemed reasonable, let’s remove and replace them with blank or nan values. One way to identify what’s reasonable is to use Interquartile range (IQR), but for now let’s use the values you’ve determined.

Once you’ve removed these values, let’s replace all the blank or null values with the mean value of the column that you calculated already.

Now that we’ve dealt with the missing and outlier values in this column, let’s recalculate the summary statistics (as a reminder this is minimum, maximum, mean, median, and standard deviation). Input these values in the ***Post-cleaning Values*** table in 2g.

What do you notice about the changes between the mean pre- and post-cleaning versus the median between pre- and post-cleaning?

### 2b. Blood Oxygen Levels

Moving on the blood oxygen column, let’s first calculate the summary statistics before we do any cleaning on the data (remember to calculate these values from the original table). Input these values in the ***Pre-cleaning Values*** table in 2g.

Now let’s go back to the copied table that we’ve started to clean. Observe the data and answer the following questions: How many missing values do we have? What is a normal range for this variable?

Since this variable is measured as a percentage, anything over 100 is an error. For any value over 100 remove those values replacing them with a blank (excel) or nan (python).

Recalculate the summary statistics for this variable and input these values in the ***post-cleaning Values*** table in 2g.

### 2c. Step Count

For the step count column, first calculate the summary statistics and input these values in the ***Pre-cleaning Values*** table in 2g.

For this variable, look at the range of values – are there any outliers? You now get to decide how you’d like to deal with the outlier and missing values (remove/leave them, impute them, etc.) Explain your decision below:

Recalculate the summary statistics for this variable and input these values in the ***post-cleaning Values*** table in 2g.

### 2d. Sleep Duration

For the sleep duration column, first calculate the summary statistics and input these values in the ***Pre-cleaning Values*** table in 2g.

For this variable, look at the range of values – are there any outliers? You now get to decide how you’d like to deal with the outlier and missing values (remove/leave them, impute them, etc.) Explain your decision below:

Recalculate the summary statistics for this variable and input these values in the ***post-cleaning Values*** table in 2g.

### 2e. Activity Level

Now let’s move over to some text data – Activity Level. First let’s identify the unique values in this column, list them below:

Do any of these appear like they’re encoding the same response (e.g., ‘High’ vs. ‘High Activity’)? Let’s fix these values in our ongoing cleaning table. Describe how you want to do this:

### 2f. Stress Level

Finally, let’s look at our last column Stress Level. First, calculate the summary statistics and input these values in the ***Pre-cleaning Values*** table in 2g.

Beyond the numerical values, what other values do you notice in the stress level column that participants self-reported?

What do you think this value means? And do you think you could “impute” a replacement value for this self-reported value?

If so, go ahead and make those changes, and recalculate the summary statistics inputting these values in the ***post-cleaning Values*** table in 2g.

### 2g. Summary Table

Input the summary statistics you’ve calculated above in the table here to summarize the variables in the dataset.

Pre-cleaning Values

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Mean | Median | Maximum | Standard Deviation |
| Heart Rate |  |  |  |  |  |
| Blood O2 Level |  |  |  |  |  |
| Step Count |  |  |  |  |  |
| Sleep Duration |  |  |  |  |  |
| Stress Level |  |  |  |  |  |

Post-cleaning Values

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Mean | Median | Maximum | Standard Deviation |
| Heart Rate |  |  |  |  |  |
| Blood O2 Level |  |  |  |  |  |
| Step Count |  |  |  |  |  |
| Sleep Duration |  |  |  |  |  |
| Stress Level |  |  |  |  |  |

## Data Analysis

Now that you’ve explored, cleaned, and summarized your data, it’s time to analyze it. The goal in this section is to test your research questions and interpret your findings clearly and thoughtfully.

We can return to our two questions:

1. Is there a significant difference in the number of steps taken across self-reported activity levels among participants?
2. What is the correlation between sleep duration and user reported stress levels?

### 3a. Does cleaning make a difference?

To make sure we can adequately communicate our findings and our decisions, let’s see if the cleaning methods we picked make any difference to the distribution of values. Create two scatterplots of Heart Rate vs. Blood Oxygen Level: one using the original data table and one using the cleaned data table. Past those figures here:

What do you notice about the two figures? What do you think caused this effect?

### 3b. Comparing Step Count Across Activity Levels

Let’s revisit the first research question: Is there a significant difference in the number of steps taken across self-reported activity levels among participants?

First let’s visualize this comparison. Let’s create either a histogram or a box plot for each activity level to compare the three groups. Paste your figure here:

Interpret your results and summarize your findings in 2-3 sentences as if you were explaining them to your manager at Steppr (who is not a data analyst).

### 3c. Relationship Between Sleep Duration and Stress Level

Your second research question asks: What is the correlation between sleep duration and user reported stress levels?

Again, let’s start with a visualization. Create a scatter plot with Sleep Duration on the x-axis and Stress Level on the y-axis. Paste your figure here:

Let’s interpret your results:

1. Look at the overall shape. Does it seem like people who sleep more tend to report more stress or less stress?
2. Calculate the correlation coefficient between Sleep Duration and Stress Level. What is the value and how might you interpret it?

Note: A number near 1 is a strong positive relationship, a number near 0 is little to no relationship and a number near -1 is a strong negative relationship.

1. Your supervisor at Steppr interpret this as “more sleep causes less stress”, is this true? How might you explain your findings to your supervisor?

Congratulations! You've completed data cleaning on this dataset.

Make sure you submit your final edited table, this completed word document, and any coding documents that you may have generated.