**Basic Data Frame:**

We have looked at the basic structure of the dataset on AM, PM and the END OF DAY daily data in a 3 month dataset. In particular, we worked on the AM and PM data and thoroughly analyzed these datasets.

In the AM dataset, we noticed that there were several columns filled with only 0s. These columns were RESPIRATION, BODY\_BATTERY, STEPS, CALORIES, FLOORS, INTENSITY\_MINUTES. As these would provide no additional information when performing data analysis, it makes sense to drop these columns for future analysis.

In addition to these columns being filled with 0s, we noticed that the DISCOMFORT\_SLIDER was almost filled with only 1s (aside from 3 rows which had values 2,3, and 4). Thus, this column also may not provide us with too much information as the majority of data is exactly the same aside from a very small chunk. We noticed that the heart rate data and stress data which had valuable information had many rows with missing values. Thus, it is essential to figure out a way to handle these missing values whether it be dropping them, replacing them with zero or other methods.

**DATA LIMITATIONS AND POTENTIAL RISKS:**

For starters, due to the various columns simply only being values of 0, they provide no new information. As a result, it makes sense to simply remove these columns. However, this simply means that we have less covariates to use for analysis. This is one potential limitation.

Another limitation is the NaN values in the dataset. The more NaN values there are in a dataset, the less information we have which leads to predictions that are less accurate. There is a big risk with handling NaN values because we need to do it in the correct way. If we do it incorrectly, it is possible to reach conclusions that are not true due to skewing the data. For preliminary analysis, we have substituted all the NaN values with the value of 0. However, this is a naive approach and will be further improved as we work through these datasets.

Another risk is to handle the data in a safe way. Due to the sheer number of different data sources we have, it is very difficult to conclude that our predictions are indeed correct. Hence, we need to ensure to take safe precautions to understand what the code is doing as it can be lost due to the sheer volume of data.

**DATA CLEANING:**

As mentioned above, we replaced all the NaN values with 0. This is a naive approach and will be changed in the future in various different ways such as using the average instead or simply removing these rows.

**DATA CORRELATION MATRIX:**

We performed a heatmap matrix that describes the Am data set by displaying the correlation between each data column. And we found out that the following column has the most significant correlation: Longitude vs latitude and Heart\_rate vs Stress.

For longitude and latitude, the correlation comes to 0.95, which is very significant. However since it is location data, it does not give too much information we were looking for. Even though the correlation between Heart\_rate and Stress is around 0.52, we can do a linear regression analysis.

**LINEAR REGRESSION ASSUMPTIONS EXAMINATION:**

To examine the assumptions of doing a linear regression, we have to check the following assumptions are fully met:

1. The data needs to be normally distributed
2. The standard deviation is randomly distributed.

If the two assumptions are met, we can say that the linear regression model is valid.

To begin with, we draw the QQ plot and Residual plot to see whether the raw data satisfied the assumption. However, they are not satisfied which means that we remove the data located at value 0.

Therefore, we delete the data at value 0, and plot the QQ plot and residual plot. Here, the data follows the assumption and we can finally make linear models between heart rate and stress.

**LINEAR REGRESSION MODEL AND INTERPRETATION:**

After performing the regression, we found there exists a positive linear relationship between the HEART RATE and STRESS. And we also calculated the R^2 which is 0.51, and adjusted R^2 which is 0.5. And this means that the data can interpret 50% of the situation that we predict.

Limitations and Refined

**Key Question Preliminary Analysis:**

The key hypothesis that we were assigned to answer was “Participants who take an average of two or fewer breaks per day will report more pain and less comfort than those who take an average of four or more breaks per day ”. After conducting some preliminary data processing, it was determined that the differences between those that take more than four and those that take less than two were extremely minor. Currently, a very naive approach was taken where all the pain was summed up for each individual and this was columned as “Total Pain”. After that, the two groups were formed of 2 or fewer breaks and 4 or more breaks and we simply averaged the total pain for every individual in each group. The “Total Pain” for two or fewer was recorded as 1.37 whereas the “Total Pain” for four or more was recorded as 1.54. While the increase in total pain is there, it seems negligible and further analysis will continue in later iterations of the deliverable. In addition, KMeans was run to see if the pain factors could predict the number of breaks by clustering. Out of the 2,308,880 pairs formed there were 574189 that disagreed which simply shows that KMeans was relatively successful in forming the clusters. In the future, we can possibly use K Nearest Neighbors and also use some logistical regression to accurately cluster and predict the data.

**Project Scope**

Our scope has changed to do the following. For starters, we will find more clever ways to see if we can use the correlation between heart rate and stress to actually predict some meaningful predictions. However, more importantly, we will focus on figuring out a way to get better results for our hypothesis. Our primary goal is to find a way to answer the hypothesis and not only that but also potentially find a way to predict the pain people are feeling with all of the data, including the number of breaks.