## Part I

1. Quick Overview
2. Get a quick understanding of the paper
3. Take a look at the data. See if you can understand what each column and row stands for.
4. Explore the data by plotting some early graphs

**Note: First send sample figures, then go ahead and plot for all**

**Note: Put the result of each sub-task (a-d) in a different folder.**

1. Per patient: All her protein/feature value (in y-axis) per period/week (x-axis).
2. Per protein: All the values of that protein for all the patients (in y-axis) per period/week (x-axis)
3. Per protein: All the values of that protein for all the patients (in y-axis) per period/week (x-axis). Additionally, differentiate patients with normal pregnancy and the rest using different colors (It is indicated as 1 or a 0 in column LatePE).
4. Do all of task “2.c”. Only difference is that instead of using the absolute value, normalize the values (using a python/R function) such that they are in the range of 0-1
5. Summary: Provide a summary of what you can observe about the trends by looking at the graphs
6. Explore the data by plotting some early graphs (Currently, in x-axis, you plot the exact week where the protein data was collected. Now, you have to perform binning. All the weeks that fit in period 1 (see paper, I suppose week 8-12 is one period, i.e., period-1) have to be put together as one point in x-axis.)

**Note: Put the result of each sub-task in a different folder.**

1. Do 2.a with one change. Instead of using the exact week in the x-axis, have the following legends: p1, p2, p3, p4, p5. All the weeks that fit in p1 should be grouped to that bin and so on.
2. Do the same with 2.b, 2.c, 2.d
3. Summary: Provide a summary of what you can observe about the trends by looking at the graphs

## Part II

1. Evaluations per period

NOTE: Do this for each period

1. Split data into training and test. Use random forest on the training data and then check how the test data performs. Use area under the curve to study the results.
2. Provide a summary of what you observe.
3. Now, try to study which proteins are contributing the most to obtaining good results. Basically, try to identify the the top 5, 10, 15, 20, .. proteins that are sufficient to provide close enough results. You can take a look at the variable called “feature importance” in python scikit or a similar function in “R”and also at “recursive Feature Elimination(RFE)” methods to study this. Provide a summary of what you observe for each period and overall.

## Part III

1. Perform a time-series analysis of the data based on plots performed in 2.
   1. Time series analysis on a per patient basis
   2. Time series analysis on a per protein basis
2. Repeat 5 for plots performed in 3

## Part IV

1. Looking at other datasets and papers
   1. Take a look at this paper: [https://www.nature.com/articles/sdata201615#t1](https://www.nature.com/articles/sdata201615" \l "t1)
   2. Take a look at the datasets presented at the end of this paper
      1. <https://www.ebi.ac.uk/pride/archive/projects/PXD002870>
      2. [https://www.synapse.org/#!Synapse:syn2289125/wiki/232688](https://www.synapse.org/" \l "!Synapse:syn2289125/wiki/232688)
      3. <https://figshare.com/collections/A_large_dataset_of_protein_dynamics_in_the_mammalian_heart_proteome/2171334>
      4. [https://www.synapse.org/#!Synapse:syn4847184.2](https://www.synapse.org/" \l "!Synapse:syn4847184.2)
   3. Provide a report about the dataset. Basically, is there any dataset that is similar to our dataset that we could use for longitudinal study.