**Homework #2 (Convergence Informatics)**

**[Input Your ID # & name]**

|  |
| --- |
| **[Instruction]**  **(1)** If you **download** the compressed file from Blackboard and release it, **there are 2 data files (HW2\_Data1.tsv and HW2\_Data2.xlsx)**.  **(2)** **Import downloaded data into the RStudio working environment**. There are many ways, and I hope you get used to the different ways you can import data. Learn how to import data by referring to Googling or our course textbooks, such as https://rafalab.github.io/dsbook/importing-data.html. The final exam is also a practical exam, so you must study how to import data.  **(3)** Try to resolve the problems below in this state.  \* You can consult with a friend, but copying the R code is absolutely prohibited. Of course, if caught, it will be treated as F grade.  **(4)** Please fill in the blanks on the bottom report. **Compress and submit** **this report and the entire R-code (.R file)** you wrote.  **\*\*\*The R-code should be executable when the TA runs.**  **\*\*\*The submitted compressed file (.zip) must be named HW2\_YourID\_Yourname.zip.** |

**Due date: November 30th 11:59PM**

**[Q1 – Q3] is based on “*HW2\_Data1.tsv*” data.**

**[Q1] After performing PCA analysis on “HW2\_Data1.tsv” data, try plotting using PC1 and PC2, respectively. When drawing a PCA plot, fill different colors depending on the disease. (Tip) There are 3 samples for each disease.**

|  |
| --- |
| **[PCA plot]** |

**[Q2] Perform clustering analysis using hierarchical clustering. Let's find the parameter in the format most similar to the PCA result (Distance Metric, Linkage Methods, and etc.) On the Dendrogram, cut and mark the optimal number of clusters you think.**

|  |
| --- |
| **[Dendrogram]** |

**[Q3] Perform clustering analysis using kmeans clustering based on Euclidean Distance. Please find optimal number of clusters (k) by changing the from 2 to 10 based on the Silhouette Score. Draw a plot for silhouette score.**

|  |
| --- |
| **[Silhouette Plot based on the kmeans clustering]** |

**[Q4 – Q6] is based on “*HW2\_Data2.xlsx*” data.**

**[Q4]** This is a record of weight loss during the summer vacation of 2021. **Let's perform predictive modeling of professor and student weight changes, respectively**. What is the best model established from professor and student data, respectively? Also, through inferred statistical comparison, **which of the two had the faster weight loss effect**? (\*Must be interpreted through **comparison of estimated statistics**)

|  |
| --- |
| **[Established model for Professor]**  **[Established model for Student]**  **[Interpretation of estimated models]** |

**[Q5]** If they continued to diet from that time until today (2021-11-22), **what is their predicted weight**?

|  |
| --- |
| **[Predicted Professor’s weight in 2021-11-22]**  **[Predicted Student’s weight in 2021-11-22]** |

**[Q6]** Let’s Visualize the given data through Scatter plots and by representing the best models estimated by professor and students, respectively, in different colors.

(\*Hint: https://rafalab.github.io/dsbook/regression.html)

|  |
| --- |
| **[Scatter plot with fitted models]** |