Coursework Part 2 – Differential Gene Expression Analysis

**Objective:**

In this coursework, you are tasked with writing a script in either Python or R to process the output of a DESeq2 differential expression analysis. The input will be a TSV file containing the results from different experimental conditions. Your goal is to output a detailed report that includes a variety of statistical analyses, visualizations, and summaries of the results.

To find your data, look in the datasets folder and take the set corresponding to your name.

**Information about the data you will use:**

Input Data: TSV files containing the results of DESeq2 differential expression analysis for different conditions. Each file includes gene identifiers, log fold changes, p-values, and adjusted p-values.

Please note: you are not expected to run DeSeq2 for yourselves.

**Report Generation:**

The output of your script should be a comprehensive report that includes the following elements:

*Summary Statistics:*

Number of significantly upregulated and downregulated genes (based on p-value and fold change thresholds) for each comparison.

Summary of p-values and log fold changes across all genes for each comparison.

*Plots:*

Volcano plots to visualize the significance and magnitude of changes in gene expression for each comparison.

MA plot to display the relationship between the log fold change and mean expression.

Histogram of p-values to assess the distribution of statistical significance.

Heatmap of the top differentially expressed genes to illustrate gene expression patterns across the conditions.

*Significant Gene Lists:*

A table or list of significantly upregulated and downregulated genes with their corresponding fold changes, p-values, and adjusted p-values.

Additional Analyses:

Any relevant insights, such as potential biological pathways affected by differentially expressed genes or clustering of gene expression patterns.

You should consider any checks for data quality before generating plots.

**Requirements**:

Your script should be able to handle the input data, process it, and output the results in a clear, organized format. You may choose either Python or R for the analysis, but the final report should include clear and interpretable results, regardless of the programming language used. The script should take as input two tsv files and generate all plots and tables that you use in your subsequent report.

An ideal report will be well-structured, including headers, legends, and captions for each plot to ensure readability and clarity of the results.

Your code needs to run on unseen tsv files. It should also include error handling for missing values or malformed input data.

**Your github repository should include:**

1. A single R or Python script named script.R or script.py, containing the code you used to answer the questions above.
2. A text file named requirements.txt containing a list of the R packages or python modules that need to be installed for your code to run.
3. A word document named output.docx containing the plots and tables generated by your code. You may add additional commentary and explanation to these plots.

You can create a zip download of your github repository once you have completed the coursework and uploaded all your files by clicking the green ‘code’ button at the top right-hand side of your repository and then selecting ‘Download ZIP’. Please upload this zip file, along with the URL of your github repository (which can be entered as text) to the Moodle Coursework 1 submission box if you completed this challenge in R or Coursework 2 submission box if you completed it in Python.