Assignment 3: Python Data Analysis

Purpose

This assignment successfully practiced core Python data analysis techniques, demonstrating my proficiency in:

- **Data Handling:** Efficiently manipulating data with NumPy and Pandas.
- **Data Visualization:** Creating informative visualizations using Matplotlib and Seaborn.

Task

1. Data Preparation

- **Loading and Cleaning:** I used Pandas to read "Gene_Expression_Data.xlsx", "Gene_Information.csv", and "Sample_Information.tsv" into DataFrames. Careful attention was paid to data types and ensuring compatibility for merging.
- **Renaming and Merging:** Based on "Sample_Information.tsv", I updated sample names in "Gene_Expression_Data.xlsx", addressing any duplicate column issues. I then merged the datasets on appropriate keys.
- **Data Splitting:** Using phenotype labels, I separated the merged dataset into "tumor" and "normal" groups for subsequent analysis.

2. Analysis

- **Average Expression:** NumPy's mean functions were applied to calculate average probe expression within the "tumor" and "normal" datasets.
- **Fold Change:** I calculated fold changes between tumor and control expressions: ((Tumor - Control) / Control).
- **Gene Filtering and Augmentation:** Data from "Gene_Information.csv" was joined, and I filtered for genes with an absolute fold change exceeding 5. Then, I added a column labeling each gene as having 'higher' expression in "Normal" or "Tumor".

3. Exploratory Data Analysis (EDA)

- **Histograms:** Matplotlib's hist() function revealed the chromosomal distribution of differentially expressed genes (DEGs), both overall and by sample type (Normal/Tumor).
- **Bar Chart:** A Seaborn bar plot visualized the percentage of upregulated and downregulated DEGs in "Tumor" samples, offering insights into tumor-specific gene activity.

- **Heatmap and Clustermap:** I employed Seaborn's heatmap() and clustermap() to create compelling visual representations of gene expression patterns across samples. These highlighted clustering trends and potential relationships.