Assignment: Multi-omics Data Exploration and Visualization

You are provided with a dataset (Table S3), containing for multiple cancer types from a given study (Wang et al., 2022):

- **Mutation frequency** of genes
- **Copy number variation (CNV) counts** across 5 categories (-2 = deep deletion, -1 = deletion, 0 = neutral, 1 = gain, 2 = amplification)
- **Differential methylation statistics** (|logFC|, adjusted p-value)
- **Differential expression statistics** (|logFC|, adjusted p-value)

Your task is to **explore, visualize, and statistically analyze** the dataset. Treat this as if you were performing the first-pass exploratory data analysis (EDA) for a genomics project.

Tasks

1. Data Preparation

- Load the dataset and check its structure.
- Perform basic summary statistics, pan-cancer and across each cancer type:
    - Distribution of mutation frequencies across all genes.
    - Distribution of CNV events (CNV_-2 ... CNV_2).
    - Distribution of methylation and expression |logFC| values.
    - Derive inferences if any.

2. CNV Analysis

- Identify the **top 10 genes** with the highest number of amplifications (CNV_2) and deletions (CNV_-2) in each cancer type.
- Visualize CNV distributions for these genes.
- Draw inferences in terms of CNV markers in specific cancers

3. Multi-omics Integration

- Explore the relationship between **CNV and expression**:
    - Correlate CNV counts (CNV_1 + CNV_2 minus CNV_-1 + CNV_-2) with differential expression (De_|logFC|).
    - Show scatterplots and correlation statistics.
- Explore the relationship between **methylation and expression**:
    - Correlate Me_|logFC| with De_|logFC|.

- Highlight genes with significant changes in both (adjusted p-value < 0.05 for methylation and expression).

4. Statistical Testing

- Perform a test (t-test, Wilcoxon, or suitable alternative) to evaluate whether **genes with amplifications (CNV_2 > 0)** show significantly higher expression logFC than genes without.
- Similarly, test whether **genes with deletions (CNV_-2 > 0)** show significantly lower expression logFC.

5. Visualization Deliverables

Produce at least the following plots (Python/Matplotlib/Seaborn or R/ggplot2):

- Histogram of mutation frequencies.
- Heatmap of CNV values for the top 50 most variable genes.
- Volcano plots for methylation and expression results (|logFC| vs -log10 adjusted p-value).
- Scatterplot of methylation vs expression |logFC|.

6. Reporting

- Write a short **report (2–3 pages in Markdown)** summarizing:
  - Key observations from your analysis
  - Notable genes where multiple alterations coincide (e.g., CNV + methylation + expression change)
  - Any interesting cancer-type specific patterns if you extend to subsets of the data

Deliverables

1. **Code** (well-commented; Python R notebook/script).
2. **Plots** as requested above.
3. **Short report** with interpretations (Markdown in notebook or PDF).