**About Dataset –**

This is a protein data set retrieved from Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB).

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB. There are two data files:

* **pdb\_data\_no\_dups.csv** contains protein meta data which includes details on protein classification, extraction methods, etc.
* **data\_seq.csv** contains >400,000 protein structure sequences.

**The `pdb\_data\_no\_dups.csv` file contains the following columns:**

1. **`structureId`**: An identifier for the structure, usually a unique code used in the Protein Data Bank.

2. **`classification`**: The biological classification of the macromolecule (e.g., DNA, protein, etc.).

3. **`experimentalTechnique`**: The technique used to determine the structure, such as X-ray diffraction.

4. **`macromoleculeType`**: The type of macromolecule described by the entry (e.g., protein, DNA/RNA Hybrid).

5. **`residueCount`**: The number of residues in the structure, which can correspond to amino acids in a protein or nucleotides in DNA/RNA.

6. **`resolution`**: The resolution at which the structure was determined, measured in Angstroms (Å). A lower number indicates a higher resolution.

7. **`structureMolecularWeight`**: The molecular weight of the structure measured in Daltons.

8. **`crystallizationMethod`**: The method used for crystallizing the macromolecule, which is critical for X-ray crystallography.

9. **`crystallizationTempK`**: The temperature at which crystallization was performed, measured in Kelvin.

10. **`densityMatthews`**: Matthew’s coefficient, a value that is used to estimate the solvent content of a crystal and can suggest the quality of the crystal structure.

11. **`densityPercentSol`**: The percentage of the crystal that is solvent, as opposed to macromolecule.

12. **`pdbxDetails`**: Additional details about the structure determination, which can include specific conditions or reagents used in crystallization.

13. **`phValue`**: The pH value at which crystallization was attempted, which can influence the structure and stability of the macromolecule.

14. **`publicationYear`**: The year in which the structure was published.

**The `pdb\_data\_seq.csv` file contains the following columns:**

1. **`structureId`**: The unique identifier for each structure entry in the Protein Data Bank (PDB).

2. **`chainId`**: The identifier for a specific chain within the structure. Proteins can have multiple chains, which are typically differentiated by letters (A, B, C, etc.).

3. **`sequence`**: The sequence of amino acids (for proteins) or nucleotides (for DNA/RNA) in the chain specified by `chainId`.

4. **`residueCount`**: The number of residues in the chain. For proteins, a residue corresponds to an amino acid, while for nucleic acids, it corresponds to a nucleotide.

5. **`macromoleculeType`**: The type of macromolecule that the sequence represents, such as DNA, RNA, Protein, or DNA/RNA Hybrid.

**Data Preprocessing –**