**1. Clinical Data (clinical\_data\_small.csv)**

This file contains patient-level demographic and clinical information.

**Main Columns:**

| **Column** | **Meaning** | **How You Might Use It** |
| --- | --- | --- |
| Patient\_ID | Unique patient identifier | Use to join across datasets or track individuals |
| Sample\_ID | Sample identifier (patient + tissue) | Join with mutation data |
| Age\_at\_Diagnosis | Age in years at diagnosis | Analyze age distribution by cancer type |
| Gender | Male/Female | Compare gender-based prevalence or outcomes |
| Race | Racial background | Study disparities in cancer occurrence/outcomes |
| Cancer\_Type | Type of cancer (e.g., breast, lung) | Stratify analyses |
| Subtype | More specific cancer subtype | See survival rates by subtype |
| Stage | Tumor stage (I–IV) | Survival analysis by stage |
| Grade | Tumor grade (1–3) | Compare aggressiveness |
| ER\_Status, PR\_Status, HER2\_Status | Receptor markers (important in breast cancer) | See mutation patterns by biomarker status |
| Family\_History | Yes/No | Explore genetic predisposition |
| Smoking\_Status | Never/Former/Current | Risk factor analysis |
| Treatment\_Type | Type of treatment received | See which treatments improve survival |
| Survival\_Time\_days | Survival time from diagnosis | Core for Kaplan–Meier survival plots |
| Vital\_Status | Alive/Deceased | Binary outcome for survival models |

**2. Mutation Data (mutation\_data\_small.csv)**

This file contains the genomic variants found in each patient sample.

**Main Columns:**

| **Column** | **Meaning** | **How You Might Use It** |
| --- | --- | --- |
| Hugo\_Symbol | Gene name (e.g., TP53) | Identify most frequently mutated genes |
| Entrez\_Gene\_ID | Numeric gene ID | Reference for databases |
| Chromosome | Chromosome number | Map mutations across genome |
| Start\_Position, End\_Position | Genomic coordinates | For genome visualization |
| Reference\_Allele | Original DNA base | Identify base change |
| Tumor\_Seq\_Allele | Mutated DNA base | Compare with reference |
| Variant\_Type | Mutation type (SNP, INS, DEL) | Mutation profile analysis |
| Variant\_Classification | Functional impact (missense, nonsense, etc.) | Predict functional damage |
| Protein\_Change | Protein-level mutation | Link to functional studies |
| Sample\_ID | Links to clinical data | Merge datasets |
| Mutation\_Status | Somatic or germline | Differentiate inherited vs acquired mutations |
| Allele\_Frequency | Fraction of reads with mutation | Assess clonality of mutation |

**Key Analysis Ideas:**

* **Survival Analysis:** Kaplan–Meier plots by mutation status, stage, or treatment.
* **Mutation Frequency:** Top mutated genes by cancer type.
* **Demographic Patterns:** Age and race breakdown by cancer type.
* **Biomarker Correlation:** ER/PR/HER2 status vs mutation patterns.
* **Risk Factor Analysis:** Smoking history vs mutation load.