**Exploratory Data Analysis Report**

**1. Introduction and Dataset Source**

The dataset used for this analysis was downloaded from **PharmGKB (The Pharmacogenomics Knowledgebase)**, a comprehensive resource that provides information about how genetic variation affects drug response. PharmGKB is a publicly available database that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers.

The clinical annotations dataset contains information about genetic variants, their associated genes, drug interactions, and clinical evidence supporting pharmacogenomic relationships.

**2. Dataset Overview**

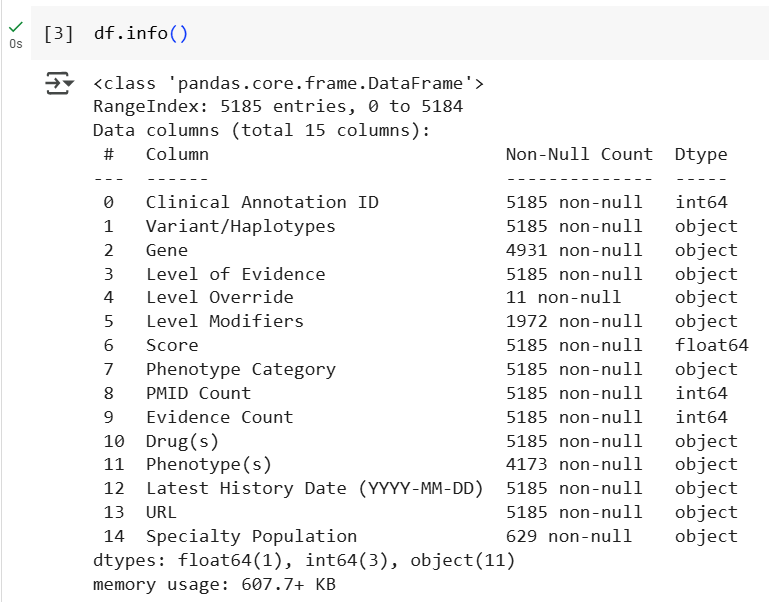
**Dataset Dimensions:** 5,185 rows × 15 columns

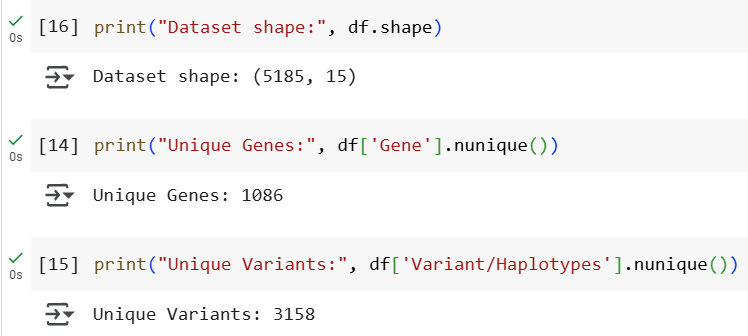
The dataset contains clinical annotations linking genetic variants to drug responses, with each row representing a unique clinical annotation entry.

**2.1 Key Columns Description:**

* **Clinical Annotation ID**: Unique identifier for each annotation
* **Variant/Haplotypes**: Genetic variants (e.g., rs951439, rs4149015)
* **Gene**: Associated genes (e.g., RGS4, SLCO1B1, CYP1A2)
* **Level of Evidence**: Clinical evidence strength (mostly level 3)
* **Score**: Numerical score indicating clinical significance
* **Phenotype Category**: Type of drug response (Efficacy, Metabolism/PK, Toxicity)
* **Drug(s)**: Associated medications
* **PMID Count**: Number of supporting publications
* **Evidence Count**: Number of evidence entries

Basic dataset information





**3. Data Quality Assessment**

**3.1 Missing Data Analysis**

The dataset shows varying levels of completeness across columns:

* **Complete data** (0 missing): Clinical Annotation ID, Variant/Haplotypes, Level of Evidence, Score, Phenotype Category, PMID Count, Evidence Count, Drug(s), Latest History Date, URL
* **Significant missing data**:
  + Level Override: 5,174 missing (99.8%)
  + Specialty Population: 4,556 missing (87.9%)
  + Level Modifiers: 3,213 missing (62.0%)
  + Phenotype(s): 1,012 missing (19.5%)
  + Gene: 254 missing (4.9%)

# Missing data analysis

**4. Dataset Characteristics and Patterns**

**4.1 Genetic Diversity**

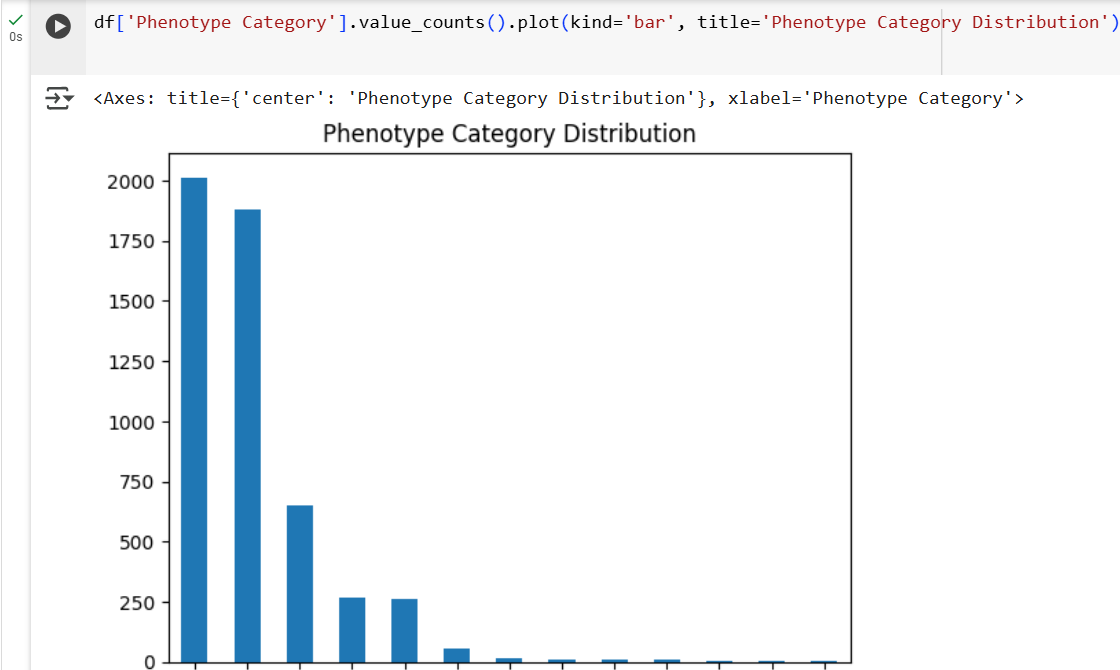
* **1,086 unique genes** represented in the dataset
* **3,158 unique variants/haplotypes**
* This indicates substantial genetic diversity in pharmacogenomic annotations

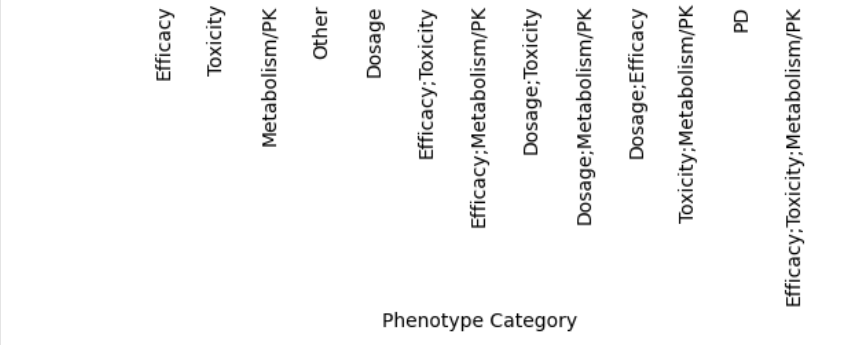
**4.2 Phenotype Categories Distribution**

The dataset focuses on three main drug response categories:

1. **Efficacy** (2,013 entries) - largest category, representing drug effectiveness
2. **Metabolism/PK** - pharmacokinetics and drug metabolism
3. **Toxicity** - adverse drug reactions and safety concerns

Phenotype category distribution

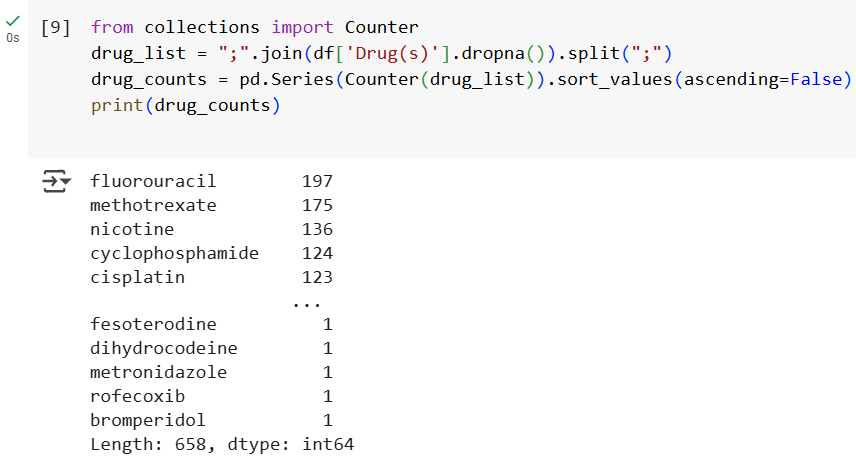




**4. Drug Coverage Analysis**

* **871 unique drugs** are represented
* **Most frequently studied drugs**:
  + Fluorouracil (197 annotations)
  + Methotrexate (175 annotations)
  + Nicotine (136 annotations)
  + Cyclophosphamide (124 annotations)
  + Cisplatin (123 annotations)

Top drugs analysis



**Clinical Evidence Strength**

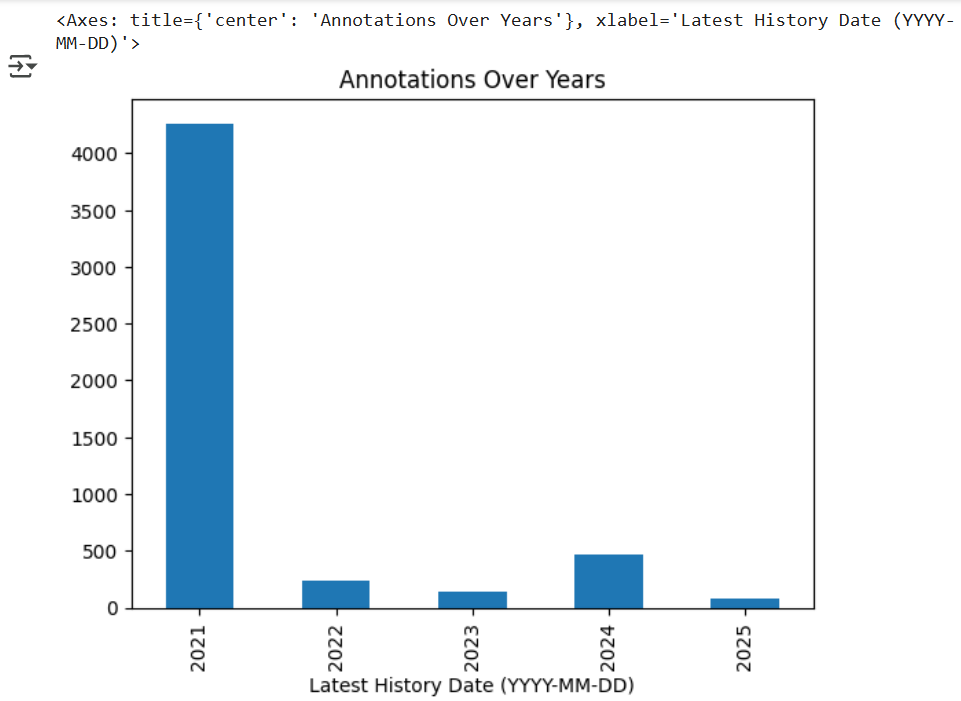
* **Score distribution**: Ranges from -37.75 to 674.25
  + Mean score: 11.28
  + Median score: 2.00
  + Most annotations have relatively low scores, indicating emerging evidence
* **PMID Count**: 1-144 supporting publications per annotation
* **Evidence Count**: 1-215 evidence entries per annotation

**5. Temporal Analysis**

The dataset shows annotations spanning multiple years, with the most recent activity in 2021 (3,891 entries on 2021-03-24), indicating active curation and updates to the pharmacogenomic knowledge base.

Temporal analysis



****

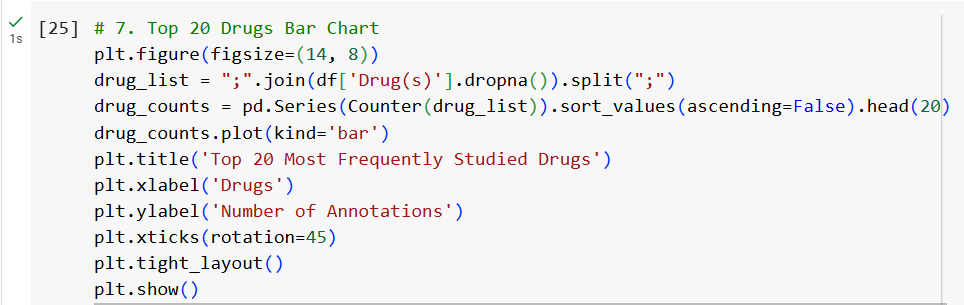
**6. Top 20 Drugs (Bar Chart)**

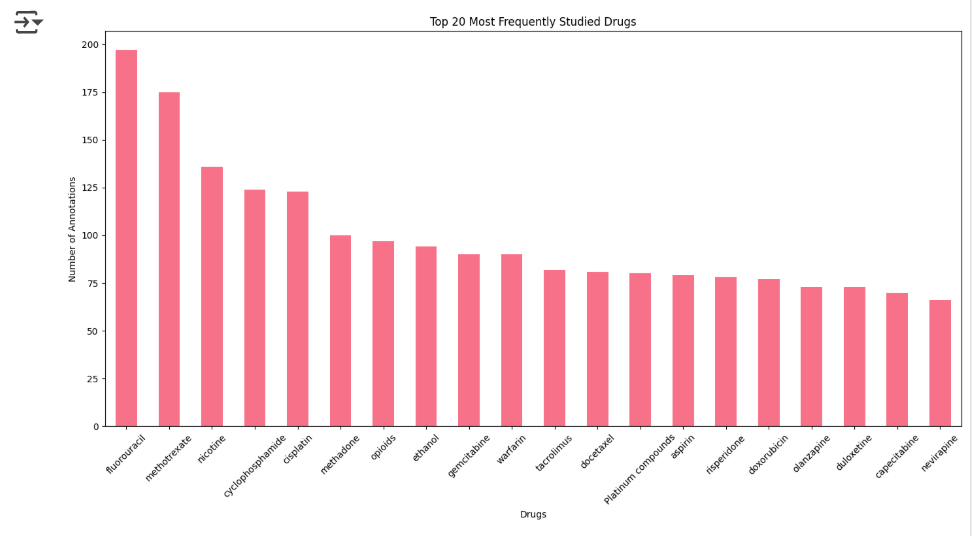
Drug frequency analysis identifies which medications have the strongest pharmacogenomic evidence base.

Key Findings:

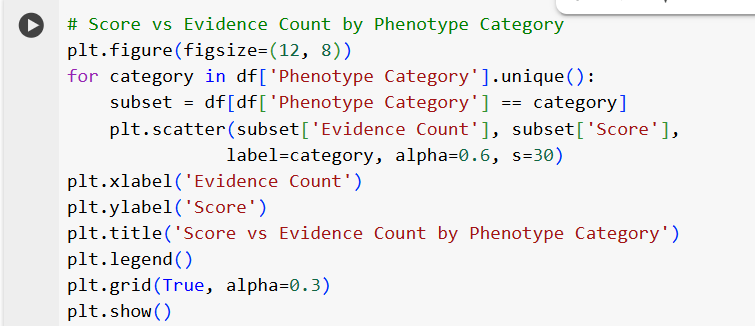
* Cancer drugs (fluorouracil, methotrexate, cyclophosphamide, cisplatin) dominate the top positions
* Psychiatric medications (antipsychotics, antidepressants) are well-represented
* The distribution reflects therapeutic areas where personalized medicine has shown clear clinical benefit

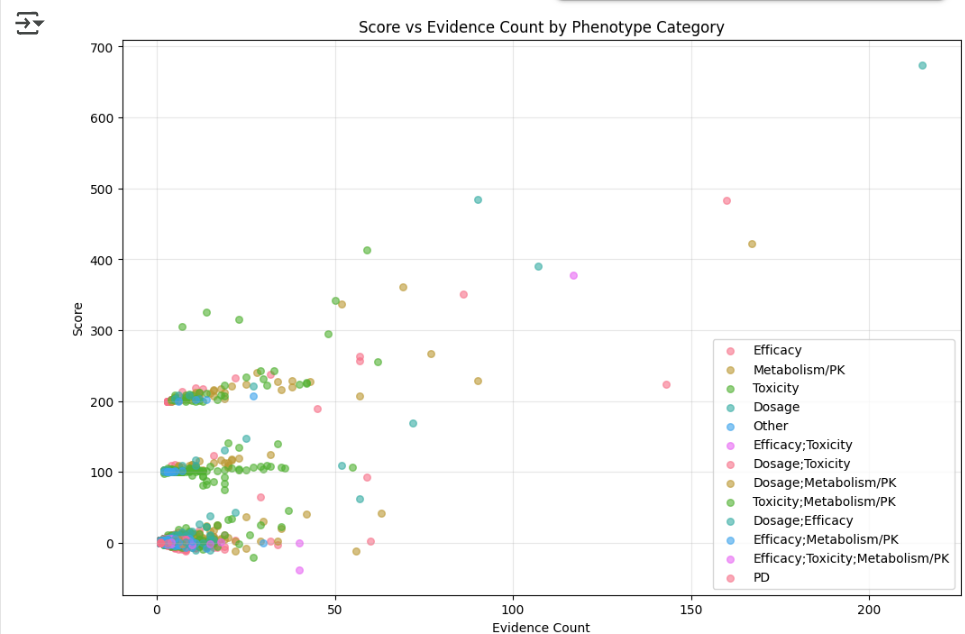
Clinical Relevance: The concentration of oncology drugs reflects the field's recognition that cancer treatment often requires personalized approaches due to genetic heterogeneity. The presence of psychiatric medications indicates growing awareness of genetic factors in mental health treatment response.

****

****

**7. Score vs Evidence Count by Phenotype Category**

****

****

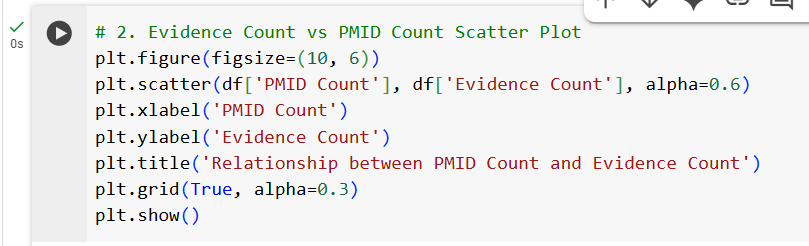
This scatter plot examines how clinical significance (Score) relates to evidence quantity across different drug response types. Most annotations cluster in the low evidence-low score region, indicating many pharmacogenomic relationships are still in early validation stages.

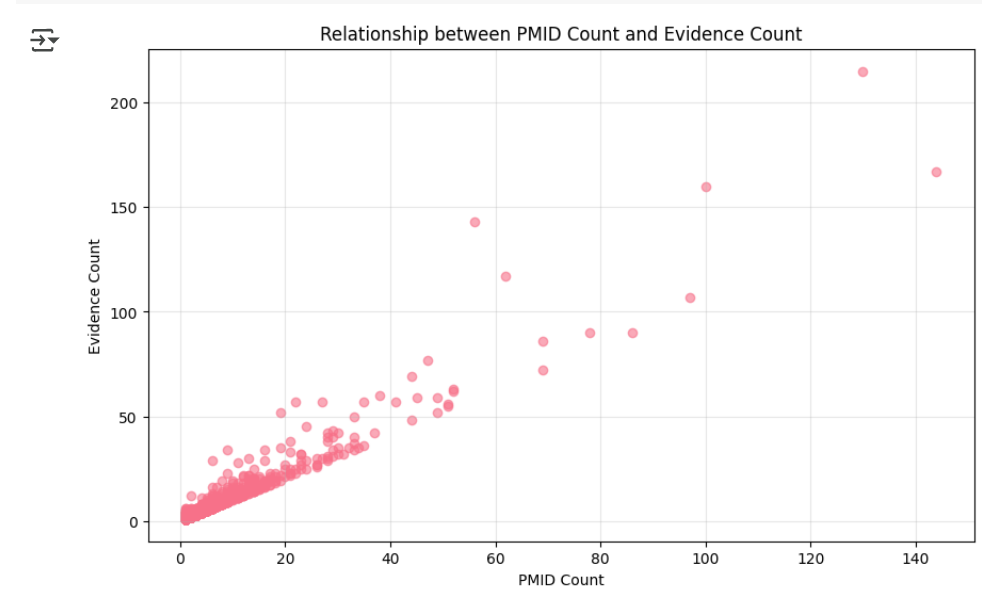
Key Findings:

* Toxicity annotations show the widest score range, with some high-impact findings despite limited evidence
* Efficacy annotations tend to achieve higher scores with moderate evidence counts
* The relationship between evidence count and score is not strictly linear, suggesting clinical significance depends on evidence quality rather than just quantity

Clinical Significance: This pattern indicates that some drug-gene interactions, particularly adverse reactions, can be clinically important even with fewer supporting studies, likely due to clear safety concerns that require immediate attention.

**8. Evidence Count vs PMID Count (Scatter Plot)**

****

****

The relationship between publication count (PMID) and evidence entries reveals patterns in how pharmacogenomic knowledge is documented.

**Key Patterns:**

* Strong positive correlation between PMID count and evidence count
* Most annotations cluster in the low PMID/low evidence region
* Some outliers show high evidence counts with relatively few publications

The positive correlation validates the evidence curation process, where more publications support more evidence entries. The clustering at low values suggests many pharmacogenomic relationships are supported by limited studies, indicating areas where additional research could strengthen clinical recommendations.

**9. Special Populations and Evidence Levels**

**Level Modifiers**

* **Tier 1 VIP**: 1,472 annotations (high clinical importance)
* **Rare Variant**: 500 annotations (204 + 296 combined)
* These modifiers help prioritize clinically actionable variants

**Specialty Populations**

* **629 pediatric-specific annotations** (12.1% of dataset)
* Indicates attention to age-specific pharmacogenomic considerations

**10. Key Insights and Clinical Relevance**

1. **Comprehensive Coverage**: The dataset covers a broad spectrum of pharmacogenomic relationships across multiple therapeutic areas
2. **Evidence-Based**: Strong emphasis on literature support (PMID counts) ensures clinical validity
3. **Quality Curation**: Use of evidence levels and modifiers shows sophisticated curation
4. **Clinical Focus**: Emphasis on efficacy, metabolism, and toxicity aligns with clinical decision-making needs
5. **Active Maintenance**: Recent update dates indicate ongoing curation efforts

**11. Potential Applications**

This dataset could be valuable for:

* Developing pharmacogenomic clinical decision support systems
* Training machine learning models for drug response prediction
* Identifying gaps in pharmacogenomic knowledge
* Supporting personalized medicine initiatives
* Conducting meta-analyses of pharmacogenomic associations

**12. Limitations and Considerations**

* High proportion of Level 3 evidence suggests many associations need stronger clinical validation
* Significant missing data in some fields may limit certain analyses
* Dataset represents current knowledge state and may evolve with new research

**13. Conclusion**

The PharmGKB clinical annotations dataset provides a comprehensive view of current pharmacogenomic knowledge, with strong emphasis on clinical evidence and practical applicability. The dataset's structure and content make it suitable for both clinical applications and research purposes, particularly in the development of personalized medicine approaches.