****Formative 1- Data quality and performance in action****

### **Part 1 – Data Quality Assessment and Improvement**

This report evaluates a dataset of **770 student records** (demographics, gaming habits, academic performance) against five core data quality dimensions: **accuracy, completeness, consistency, timeliness, and uniqueness**. The goal was to prepare the dataset for analysis by resolving structural and formatting issues.

#### **Step 1: Sex Column Standardization**

The Sex column initially used numeric codes (0 and 1). To improve interpretability, these were mapped to categorical values:

* **Formula**: =IF(A2=1,"Male",IF(A2=0,"Female","Check Value"))
* **Outcome**: Ensured consistency by replacing ambiguous codes with clear labels (Female/Male).

#### **Step 2: Percentage Column Cleaning**

The Percentage column had formatting inconsistencies (e.g., mixed symbols like "%", commas, double dots). Two formulas were applied:

1. **Error Detection**:  
   =IF(ISERROR(VALUE(SUBSTITUTE(SUBSTITUTE(K2,",","."),"%",""))), "ERROR", "OK")  
   Flagged cells with invalid formatting (e.g., "75..5%").
2. **Normalization**:  
   =IF(L2="ERROR", IF(ISNUMBER(SEARCH("..",J2)), VALUE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(J2,"..","."),",","."),"%",""))/100, "CHECK MANUALLY"), ...)  
   Converted values to decimal percentages (e.g., "75.5%" → 0.755) and flagged ambiguous entries for manual review.

#### **Step 3: Missing Value Checks**

Using regex (^$), the dataset was scanned for empty cells. **No missing values** were identified, confirming completeness post-cleaning.

#### **Limitations**

* **Accuracy/Timeliness**: Could not be fully validated due to insufficient metadata (e.g., no source documentation or timestamps).
* **Uniqueness**: Duplicates were ruled out through iterative checks.

**Outcome**: The dataset is now consistent, with standardized formats and resolved errors, making it suitable for analysis.

### **Part 2 – Database Schema Design with SQL**

A **star schema** was designed to analyze relationships between gaming habits and academic performance, prioritizing query efficiency.

#### **Schema Structure**

1. **Fact Table**: fact\_grades
   * **Fields**: grade, percentage, error\_flag (validates data integrity).
   * **Foreign Keys**: Links to all dimension tables.
2. **Dimension Tables**:
   * dim\_student: Demographics (sex).
   * dim\_school: School identifiers.
   * dim\_parent\_background: Parental education (0–10 scale) and income (1–4 scale).
   * dim\_gaming\_habits: Gaming metrics:
     + playing\_years (0–4 years),
     + playing\_often (0–5 frequency scale),
     + playing\_hours (0–5 session length).

#### **Key Design Features**

* **Constraints**: Enforced valid ranges (e.g., parental education 0–10).
* **Indexes**: Optimized join performance between fact and dimension tables.
* **Error Handling**: An unresolved formatting error in **column 365** was manually corrected post-load. This highlighted the need for stricter cleansing protocols.

#### **SQLite Implementation**

1. **Foreign Key Enforcement**: Enabled via PRAGMA foreign\_keys = ON;.
2. **Schema Creation**: Tables built with explicit data types and constraints.
3. **Data Transformation**: Cleaned data loaded into the schema using batch inserts.

### **Example Analytical Use Case**

**Query**: Compare academic performance between:

* **Frequent gamers**: playing\_often ≥ 4 (high engagement).
* **Non-gamers**: playing\_years = 0 (no engagement).

**Filters**: Parental education level (e.g., isolating students with parents scoring ≥8 on the education scale).

**Outcome**: The star schema allows efficient aggregation of grades across dimensions, enabling insights into how gaming habits correlate with academic results while controlling for socioeconomic factors.

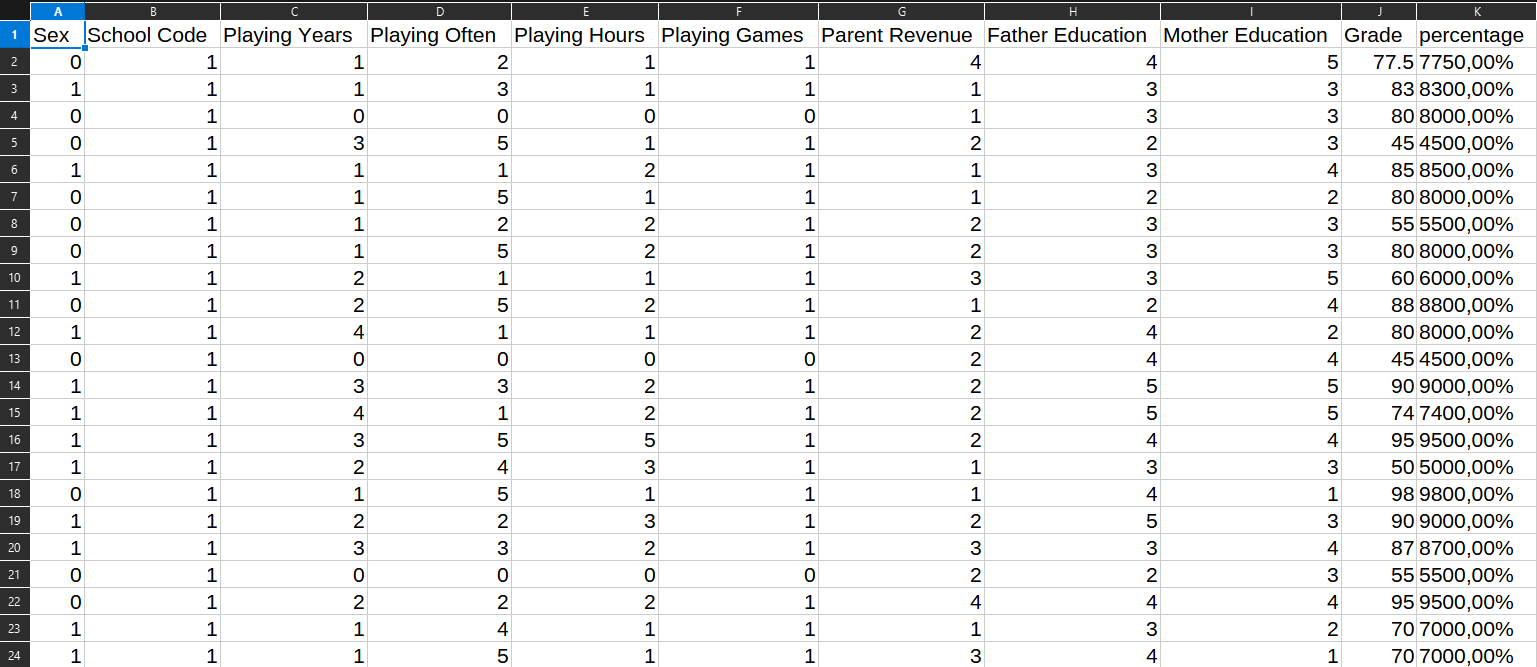
### **Conclusion**

This project underscores the importance of rigorous data cleaning and schema design in ensuring reliable analytics. While challenges like unresolved errors and metadata gaps persist, the refined dataset and optimized database structure provide a robust foundation for further analysis.

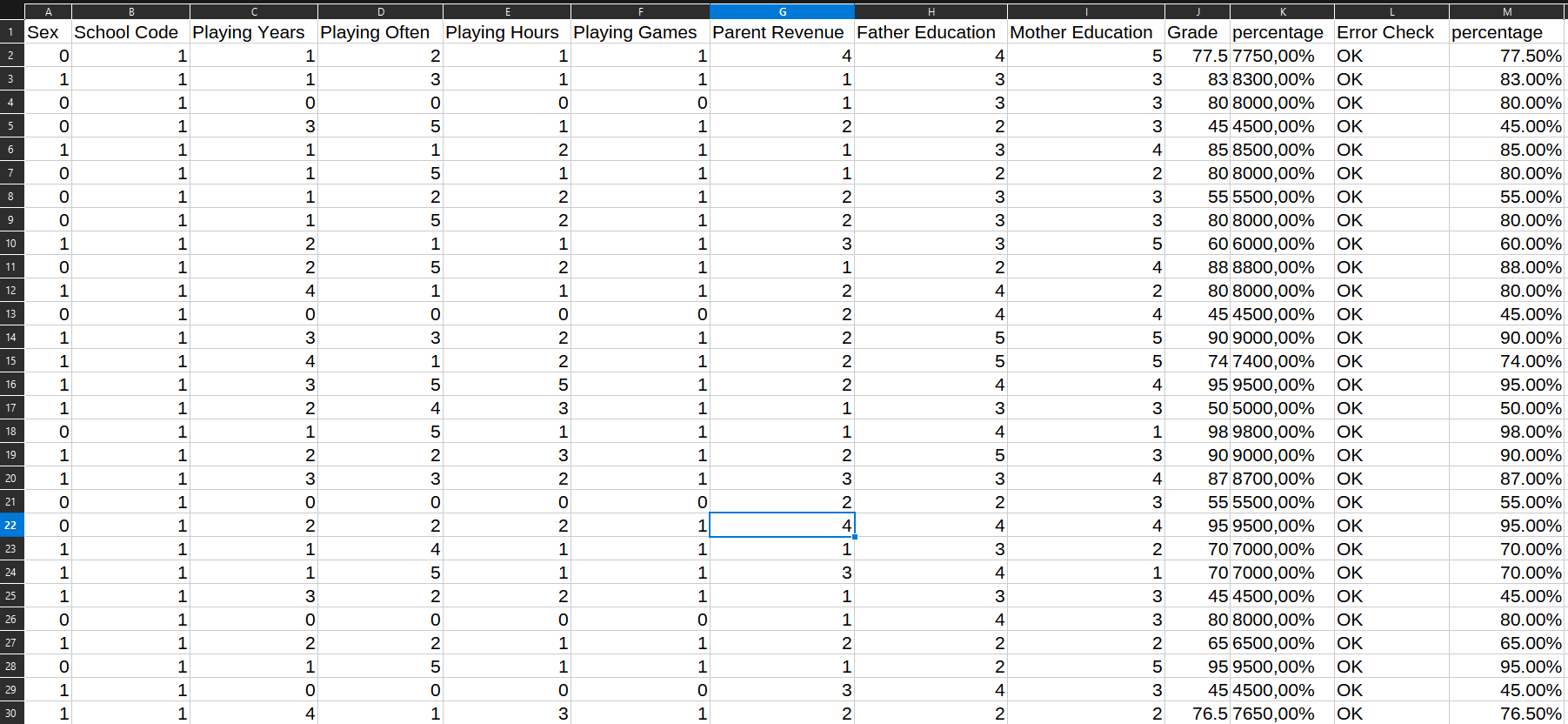
Screenshots

**Part 1**

Before:



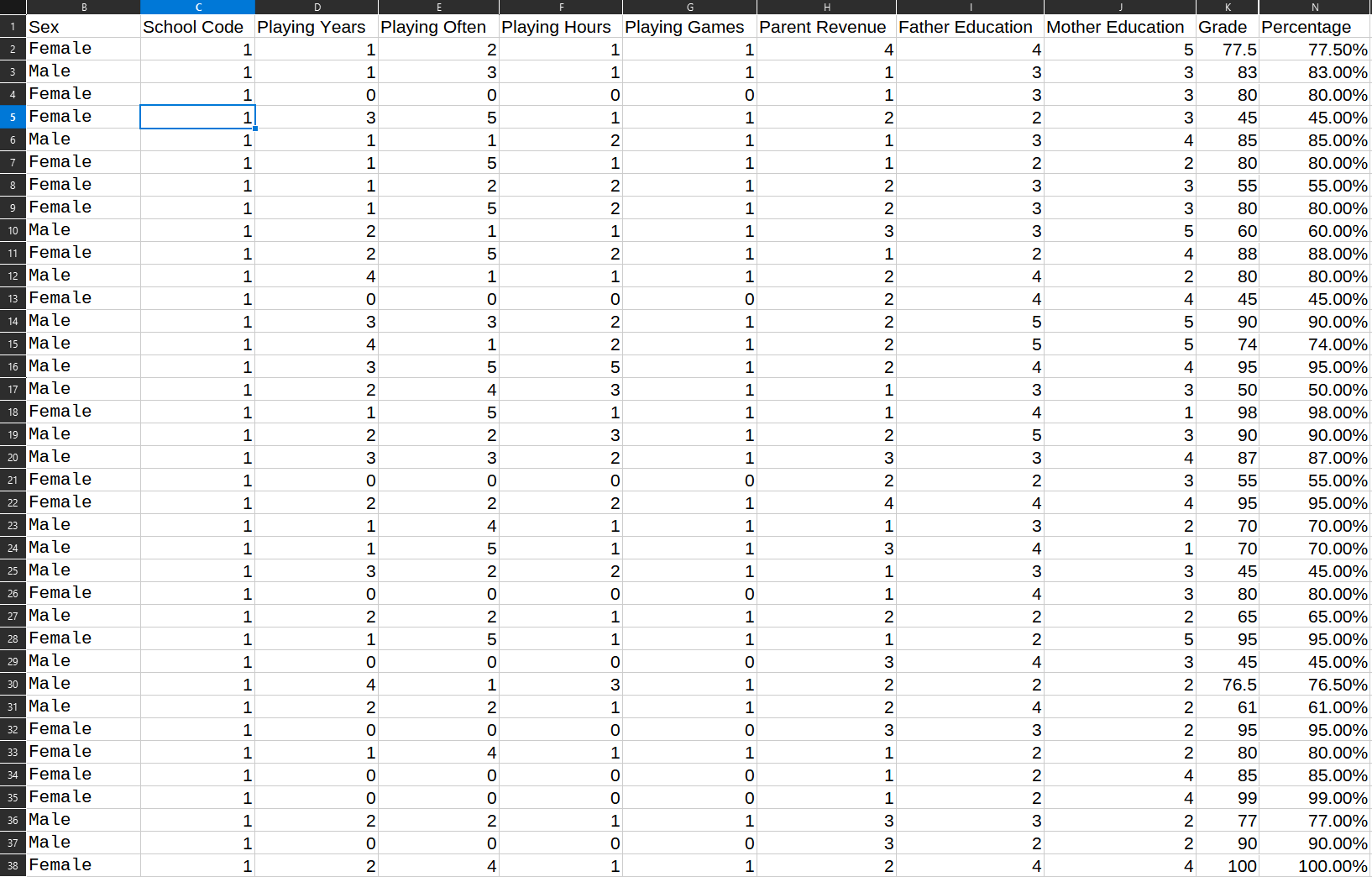
During Clean:



Error:

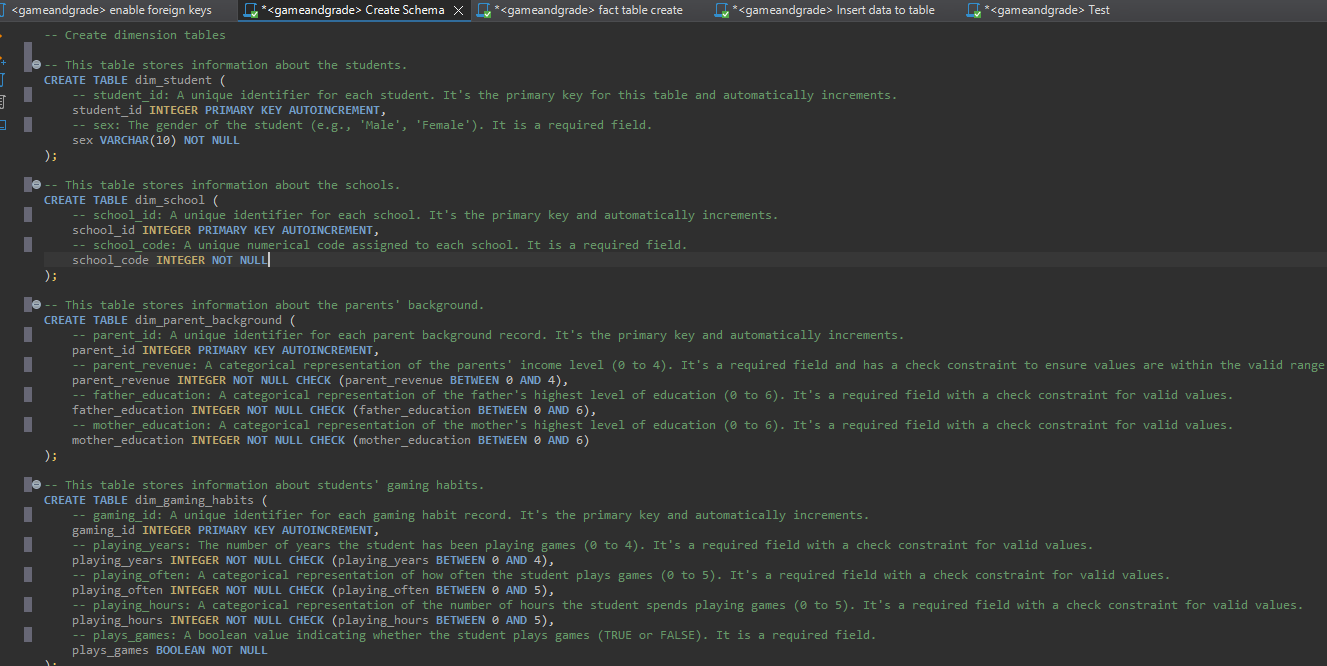


Clean:

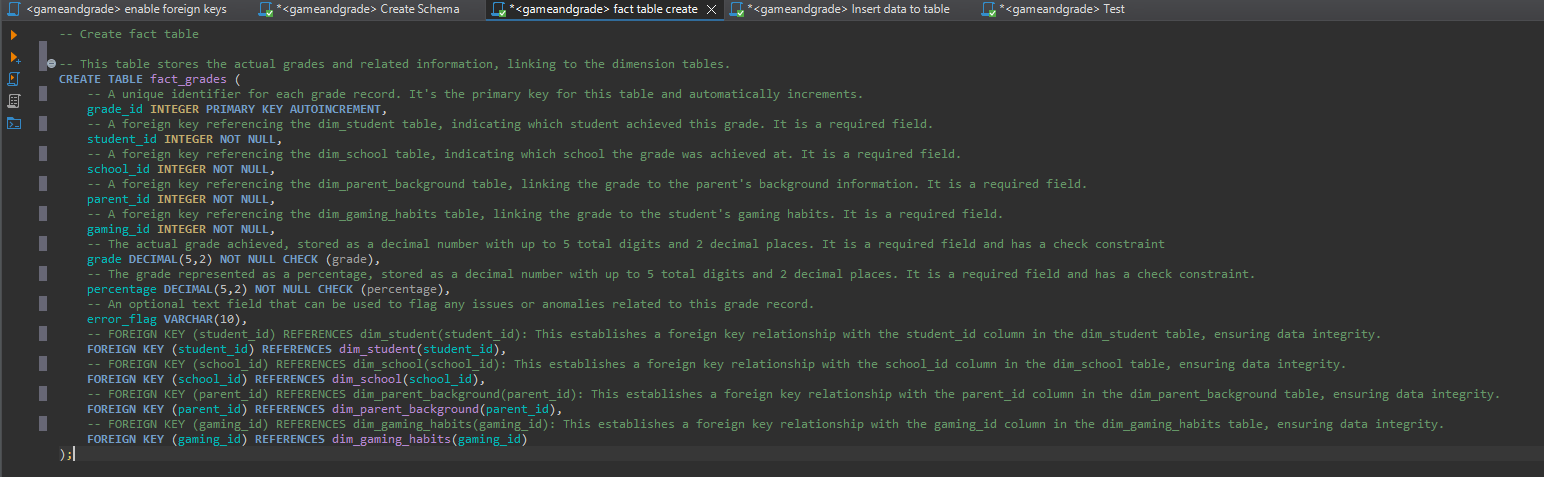


**Part 2**

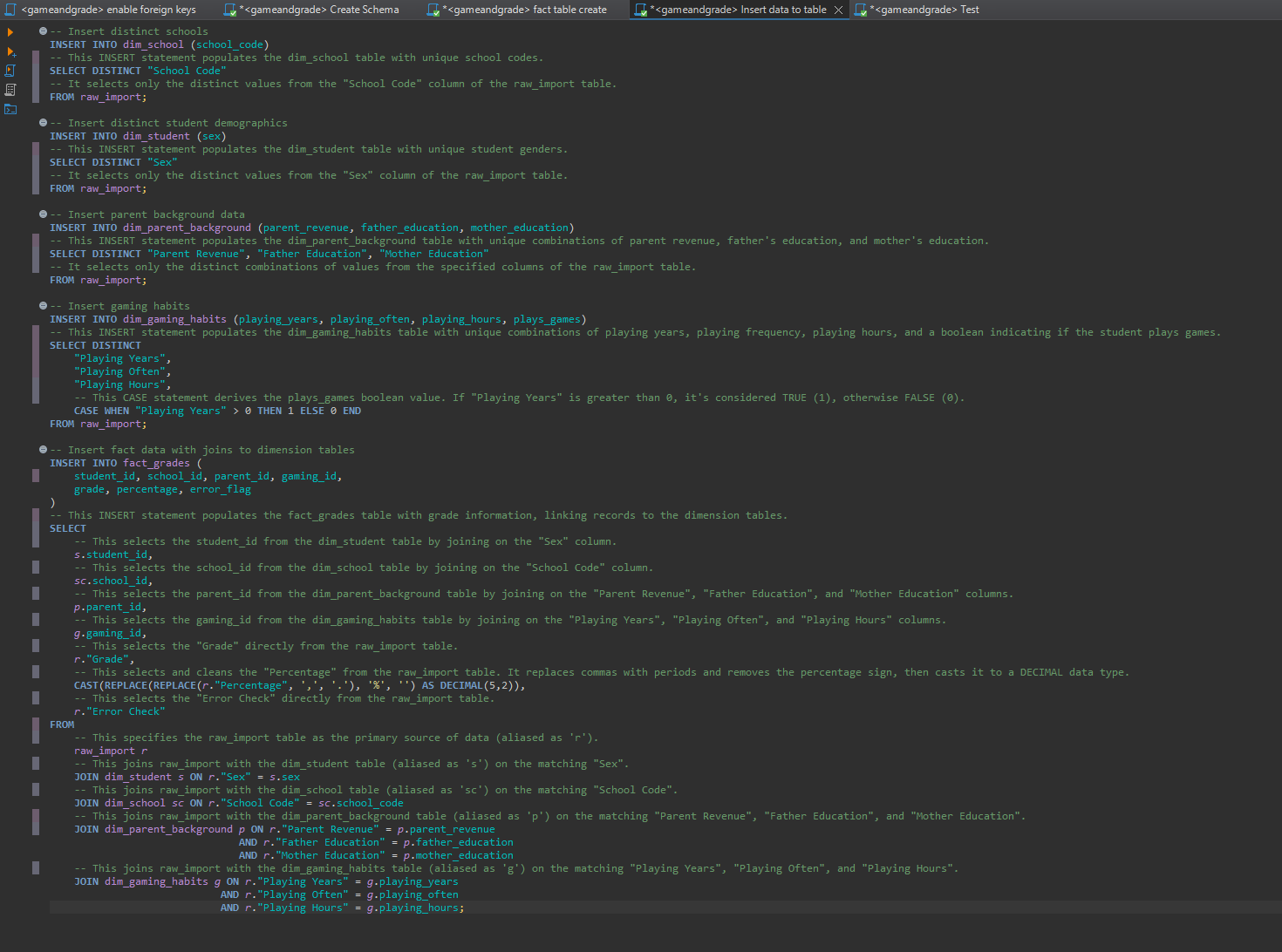
## Create Schema:



Create fact\_table:



Load and transform data:



## Example Use Case

