**Initial Data Exploration and Understanding Report**

**1. Dataset Overview**

* **File Format**: Excel (.xlsx)
* **Rows**: 1,048,575
* **Columns**: 45 (after removing unnecessary Unnamed: 0)
* **Domain**: Cybersecurity Incident Triage Classification
* **Target Variable**: IncidentGrade (with classes - **True Positive**, **Benign Positive**, **False Positive**)

**2. Feature Overview**

| **Feature Type** | **Count** | **Comments** |
| --- | --- | --- |
| Numerical (int64, float64) | ~32 | IDs, numerical metadata (e.g., OrgId, DeviceId, OSVersion, etc.) |
| Categorical (object) | ~13 | Strings like Category, MitreTechniques, Roles, ThreatFamily |
| Time-based | 1 | Timestamp (currently object, parsed later) |

**3. Target Variable (IncidentGrade) Distribution**

* **Benign Positive**: 43.3%
* **True Positive**: 35.2%
* **False Positive**: 21.5%

Slight class imbalance noticed:

* BenignPositive is the dominant class (~43%)
* FalsePositive is the minority class (~21%)
* Not heavily skewed, but imbalance **must be handled** during model training (e.g., class weights or sampling).

**4. Missing Values**

Several columns have substantial missing data:

| **Column** | **% Missing** | **Comments** |
| --- | --- | --- |
| ActionGrouped, ActionGranular | ~99% | Likely candidates for dropping unless critical |
| ThreatFamily | ~99% | Very sparse; may not be usable |
| ResourceType | ~99% | Very sparse; needs investigation |
| Roles, AntispamDirection, SuspicionLevel, LastVerdict | 80% - 97% | High missingness; requires strategic handling (either drop, impute, or use 'Unknown' category) |
| MitreTechniques | ~57% | Manageable missingness; imputation may be possible |

Columns with **no missing values** are mostly identifiers and some metadata.

**5. Duplicate Rows**

* **396 duplicate rows** found (~0.04% of dataset)
* Action: **Duplicates will be dropped** during preprocessing to avoid data leakage and bias.

**6. Timestamp Field**

* Timestamp is currently of object type.
* Parsing it into datetime format was initiated.
* From Timestamp, we can **extract new features**:
  + **Hour** (cyber incidents often have time-based patterns)
  + **Day** (incident distribution by day)
  + **Month** (seasonality patterns)

Important for potential feature engineering later.

**7. Numerical Feature Distribution**

* Most numerical fields (like OrgId, DeviceId, OSVersion) are **IDs or codes** — **not true continuous variables**.
* Their distributions are **not normally distributed**.
* Some features have **skewed distributions** — possible normalization or transformation needed later.

**8. Categorical Features**

* Some categorical features have very **high cardinality** (many unique values), especially:
  + Category
  + Roles
  + ThreatFamily
* Action: These may require **target encoding** or **embedding** rather than simple one-hot encoding.
* Categorical fields with **few unique values** can be safely one-hot encoded.

**9. Correlation Analysis**

* Numerical correlations were being computed (in progress).
* Expected: Weak direct correlations because most numerical fields are identifiers.
* Action: Focus more on categorical interaction and timestamp features rather than traditional numeric correlations.

**Summary of Key Observations**

| **Aspect** | **Observations** | **Planned Actions** |
| --- | --- | --- |
| Class Imbalance | BenignPositive > TruePositive > FalsePositive | Handle via class weights or resampling |
| Missing Values | Several features >80% missing | Drop or impute wisely |
| Duplicates | Minor duplicates found | Drop duplicates |
| Timestamp | Can be feature engineered (Hour, Day, Month) | Extract and use |
| Categorical Features | High cardinality features exist | Smart encoding required (target encoding, embeddings) |
| Numerical Fields | Mostly IDs | Treat carefully; not standard continuous features |