Exploratory Data Analysis (EDA) Summary for NSL-KDD Dataset

📊Dataset Overview

- **Total rows**: 125,973
- **Total columns**: 43

📦Structure and Features

- No missing values detected in any of the columns.
- No duplicate rows found.

🕚Feature Types

- **Categorical columns**:
- `protocol_type`: 3 unique values → ['tcp', 'udp', 'icmp']
- `service`: 70 unique values → e.g., 'http', 'smtp', 'ftp_data', 'private', etc.

- `flag`: 11 unique values → e.g., 'SF', 'S0', 'REJ', 'RSTR', etc.

- **Continuous/numeric columns**: Most of the other features (like `src_bytes`, `dst_bytes`, `duration`, etc.) are numeric and show wide range and variance.

🧮Basic Statistics (Selected Observations)

- `duration`:
- Mean: 287.14, Max: 42908, Std: 2604.5 → **highly skewed**
- `src_bytes`, `dst_bytes`:
- Values range from 0 to 1.3 billion+ → **need scaling**
- `land`, `urgent`, `wrong_fragment`, `hot`:
- Many are zero in most cases → possibly low-variance

🔍Class Distribution (`label` column)

Label	Count
normal	67,343
neptune	41,214
satan	3,633
ipsweep	3,599
portsweep	2,931
smurf	2,646

Label	Count
nmap	1,493
back	956
teardrop	892
warezclient	890
(others)	< 300 each

- **Class imbalance** is present.
- Majority class is `normal` followed by `neptune`.
- Minority classes like `perl`, `spy`, `phf` have fewer than 10 instances.

🔗Cleanliness Check

- **Missing values**: None detected
- **Duplicates**: None found
- **Ready for preprocessing**

🔍Recommendations Before Preprocessing

🕰️1. Encoding

- Encode categorical columns:
- One-hot encoding: `protocol_type`, `flag`
- Frequency or grouping: `service` (too many unique values)

🕰️2. Scaling

- Standardize or normalize high-range features:
- `src_bytes`, `dst_bytes`, `duration`

🕰️3. Feature Selection

- Consider dropping or analyzing features with:
- Little or no variance
- Redundant or constant values

🕰️4. Class Imbalance Handling

- Use techniques like:
- SMOTE / ADASYN for oversampling
- Class weighting in models
- Stratified train-test splits

📘Next Steps

1. Feature encoding (start with one-hot for `protocol_type`, `flag`)
2. Normalize selected numeric columns
3. Create label maps if needed (e.g., binary classification: normal vs attack)
4. Split into train-test sets
5. Begin model prototyping (start with a simple one like Logistic Regression)

Document generated based on preliminary EDA performed on the NSL-KDD dataset.