In [42]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import nltk
nltk.download('punkt')
import matplotlib.pyplot as plt
import matplotlib
import warnings
warnings.filterwarnings('ignore')
from wordcloud import WordCloud
import re
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [43]:
```
pip install umap-learn
```

```
Requirement already satisfied: umap-learn in /usr/local/lib/python3.10/dist-packages
(0.5.6)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages
(from umap-learn) (1.26.4)
Requirement already satisfied: scipy>=1.3.1 in /usr/local/lib/python3.10/dist-package
s (from umap-learn) (1.13.1)
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.10/dist-p
ackages (from umap-learn) (1.5.2)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packag
es (from umap-learn) (0.60.0)
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.10/dist-pac
kages (from umap-learn) (0.5.13)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from
umap-learn) (4.66.5)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.1
0/dist-packages (from numba>=0.51.2->umap-learn) (0.43.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.10/dist-package
s (from pynndescent>=0.5->umap-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.10/dist
-packages (from scikit-learn>=0.22->umap-learn) (3.5.0)
```

Read the dataset

In [44]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.m
ount("/content/drive", force_remount=True).
```

In [45]:
```
data = pd.read_csv("/content/drive/My Drive/Modified_SQL_Dataset.csv")
```

Explore information from dataset

In [46]:
```
def missing_values(data):
    """
    This is to get the percetages of missing data
    Args:
        df (pd.Dataframe): contains the data
    Returns:
```

```
        missing_percetanges(pd.Dataframe): contains Column,     Counts, and     Percer
            of the missing values for eah colmn
    """
    missing_count = data.isnull().sum()
    missing_percetanges = pd.DataFrame({
        'Column': missing_count.index,
        'Counts': missing_count.values,
        'Percentage': missing_count.values / len(data) * 100
    })
    return  missing_percetanges
```

In [47]:
```
def explore_sample(sample):
    """
    Exploring a dataset sample
    Args:
        sample (pd.Dataframe): the dataset sample to explore.
    Returns:
        results (dict): containing results of each exploration with the title as key
    """
    shape = pd.DataFrame(sample.shape)
    head = pd.DataFrame(sample.head())
    tail = pd.DataFrame(sample.tail())
    nunique = pd.DataFrame(sample.nunique(), columns=["#_of_Unique"])
    describe = pd.DataFrame(sample.describe())
    dtypes =  pd.DataFrame(sample.dtypes, columns=["Datatype"])
    results = {
        'Dataset shape:':shape,
        'Dataset Head:':head,
        'Dataset Tail:':tail,
        'Dataset Numerical Describtion: ':describe,
        'Missing Values By Percentage': missing_values(sample),
        'Dataset Columns Data types: ':dtypes,
        'Number of uniques in the datasets:':nunique}
    return results
```

In [48]:
```
def print_dataset_exploration(results):
    """
    Prints a beautufil display of each of the exploration dataframe
    Args:
        results (dict): contains exploration outputs with the title as key
    Returns:
        nothing
    """
    for operation, dataframe in results.items():
        print(f"{operation}")
        if operation == 'Missing Values By Percentage':
            print("Total Sum of Missing Percetange: ", dataframe['Percentage'].sum())
        display(dataframe)
```

In [49]:
```
results = explore_sample(data)
print_dataset_exploration(results)
```

Dataset shape:

	0
0	30919
1	2

Dataset Head:

	Query	Label
0	" or pg_sleep (__TIME__) --	1
1	create user name identified by pass123 tempora...	1
2	AND 1 = utl_inaddr.get_host_address (...	1
3	select * from users where id = '1' or @ @1 ...	1
4	select * from users where id = 1 or 1#" (...	1

Dataset Tail:

	Query	Label
30914	DELETE FROM door WHERE grow = 'small'	0
30915	DELETE FROM tomorrow	0
30916	SELECT wide (s) FROM west	0
30917	SELECT * FROM (SELECT slide FROM breath)	0
30918	SELECT TOP 3 * FROM race	0

Dataset Numerical Describtion:

	Label
count	30919.000000
mean	0.368123
std	0.482303
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

Missing Values By Percentage
Total Sum of Missing Percetange: 0.0

	Column	Counts	Percentage
0	Query	0	0.0
1	Label	0	0.0

Dataset Columns Data types:

	Datatype
Query	object
Label	int64

Number of uniques in the datasets:

	#_of_Unique
Query	30905
Label	2

Print the existing tables in the payload.

In [50]:
```
from_values = data[data.Query.str.contains('from')]
l = list(from_values.Query)

tables = []
for i in l:
    x = i[i.index('from')+5:].split(' ')
    if x[0] not in tables and len(x[0]) > 1:
        tables.append(x[0])
print(tables)
```

```
['users', 'syscolumns', 'sysobjects', 'information_schema.tables--', 'temp', 'tabnam
e', 'where', 'information_schema.tables;', 'tablenames', 'wapiti', 'generate_series',
'information_schema.character_sets', 'dual', 'pg_sleep', 'all_users', 'sysibm.systabl
es', 'rdb$database', 'sysusers', 'mysql.db', 'domain.domains', 'rdb$fields', 'maste
r..sysdatabases', 'information_schema.system_users', 'dual--', 'dual#', 'WHERE', 'RO
M', 'JOIN']
```

In [51]:
```
data.duplicated().any()
```

Out[51]:
```
True
```

Remove duplicate rows

In [52]:
```
data.drop_duplicates(inplace=True)
```

In [53]:
```
data.duplicated().any()
```

Out[53]:
```
False
```

In [54]:
```
data.duplicated(subset=['Query'],keep=False).sum()
```

Out[54]:
```
4
```

Remove same query with different label

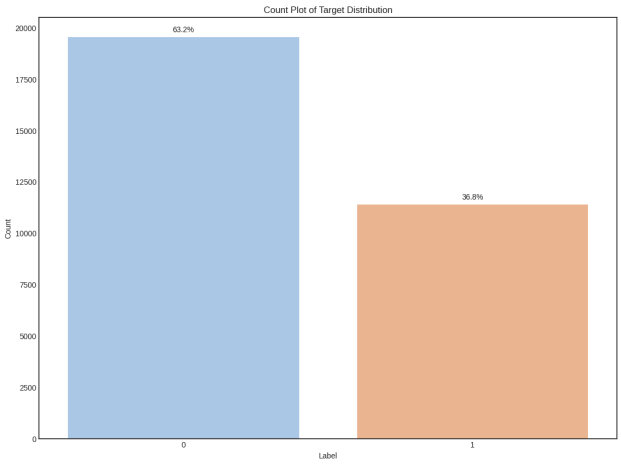
In [55]:
```
data.drop_duplicates(subset=['Query'], keep=False, inplace=True)
```

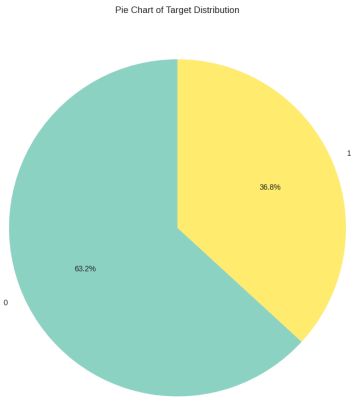
In [56]:
```
data.duplicated(subset=['Query'],keep=False).sum()
```

Out[56]:
```
0
```

In [57]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Set figure size
plt.figure(figsize=(30, 10))
```

```
# First subplot: Countplot for target distribution
plt.subplot(1, 2, 1)
ax1 = sns.countplot(x='Label', data=data,palette='pastel')
for p in ax1.patches:
    ax1.annotate('{:.1f}%'.format(100 * p.get_height() / len(data)),
                 (p.get_x() + p.get_width() / 2., p.get_height()),
                 ha='center', va='center', xytext=(0, 10),
                 textcoords='offset points')

ax1.set_title('Count Plot of Target Distribution')
ax1.set_xlabel('Label')
ax1.set_ylabel('Count')

# Second subplot: Pie chart for target distribution
plt.subplot(1, 2, 2)
ax2 = data['Label'].value_counts().plot(kind='pie', colormap='Set3', autopct='%1.1f%%'
plt.ylabel('')  # Hide y-label in pie chart
plt.title('Pie Chart of Target Distribution')

# Display the plots
plt.show()
```





How many comments symbols are in the payloads ?

In [58]:
```
comment_values = data[data.Query.str.contains('#|--|//')]
comment_values.count()[0]
```

Out[58]:
```
5925
```

Top used words in payloads

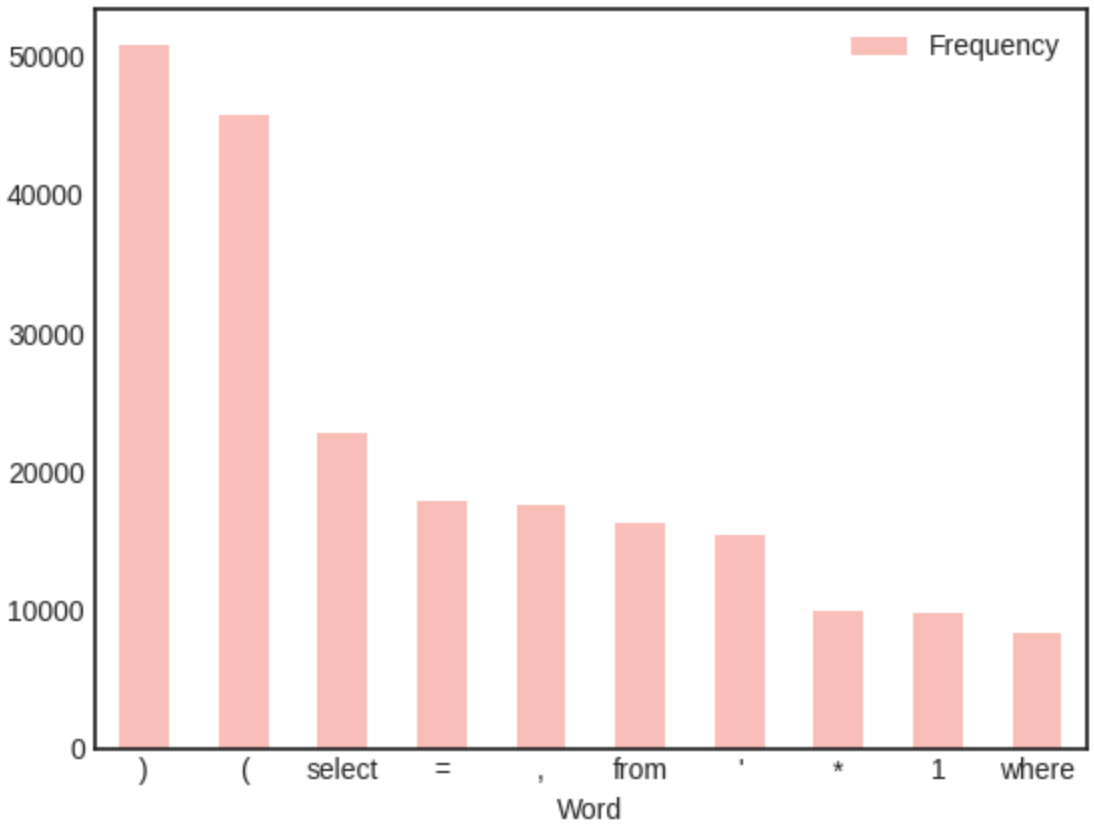
In [59]:
```
top_N = 10

payloads = data.Query.str.lower().str.replace(r'\|', ' ').str.cat(sep=' ')
words = nltk.tokenize.word_tokenize(payloads)
word_dist = nltk.FreqDist(words)


print('All frequencies, including STOPWORDS:')
print('=' * 60)
rslt = pd.DataFrame(word_dist.most_common(top_N),
                    columns=['Word', 'Frequency']).set_index('Word')
print(rslt)
print('=' * 60)
```

```
matplotlib.style.use('seaborn-v0_8-white')

rslt.plot.bar(rot=0, alpha=0.85, colormap='Pastel1')
```

```
All frequencies, including STOPWORDS:
============================================================
        Frequency
Word
)           50877
(           45822
select      22783
=           17808
,           17640
from        16342
'           15407
*            9957
1            9819
where        8357
============================================================
```

Out[59]:

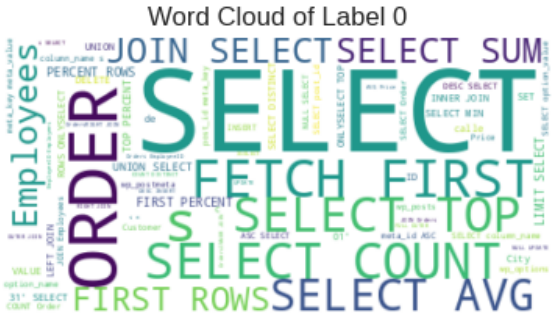
```
<Axes: xlabel='Word'>
```



In [60]:

```
data_label_0 = data[data['Label'] == 0]
words_label_0 = ' '.join(data_label_0['Query'].dropna())

data_label_1 = data[data['Label'] == 1]
words_label_1 = ' '.join(data_label_1['Query'].dropna())

wordcloud_0 = WordCloud(background_color='white').generate(words_label_0)
plt.figure(figsize=(10, 5))
plt.subplot(1, 2, 1)
plt.imshow(wordcloud_0, interpolation='bilinear')
plt.title('Word Cloud of Label 0')
plt.axis('off')
```

```
wordcloud_1 = WordCloud(background_color='white').generate(words_label_1)
plt.subplot(1, 2, 2)
plt.imshow(wordcloud_1, interpolation='bilinear')
plt.title('Word Cloud of Label 1')
plt.axis('off')

plt.show()
```



Nhận xét: Các từ khóa xuất hiện thường thuyên trong các câu lệnh SQL bình thường gồm: SELECT, ORDER, TOP, AVG, JOIN, COUNT,v.v Các từ khóa xuất hiện thường xuyên trong các câu lệnh SQL Injection gồm: null, char, chr, union, select, case, count,v.v

Feature Extraction

In [61]:
```
harmful_keywords = ['DROP', 'TRUNCATE', 'DELETE', 'SHUTDOWN','SLEEP ']
# Số dấu nháy đơn trong câu truy vấn
def single_qt(x):
    return len(re.findall(r"'", x))
data['no_sngle_quts'] = data['Query'].apply(single_qt)

# Số dấu nháy kép trong câu truy vấn
def double_qt(x):
    return len(re.findall(r'"', x))
data['no_dble_quts'] = data['Query'].apply(double_qt)

# Số lượng dấu câu trong câu truy vấn
def punctuation(x):
    r = re.compile("[!\"#$%&\'()*+,-.\/:;<=>?@[\\]^_`{|}~]")
    l = r.findall(x)
    return len(l)
data['no_punctn'] = data['Query'].apply(punctuation)

# Số dòng comment đơn trong câu truy vấn
def sin_comm(x):
    return len(re.findall(r'(--)', x))
data['no_sgle_cmnt'] = data['Query'].apply(sin_comm)

# Số lượng comment đa dòng
def mult_comm(x):
    return len(re.findall(r'(\/\*)', x))
data['no_mlt_cmnt'] = data['Query'].apply(mult_comm)

# Tổng số lượng toán tử logic trong câu truy vấn
def logical(x):
    return len(re.findall(r'\snot\s|\sand\s|\sor\s|\sxor\s|&&|\|\||!', x))
data['no_log_oprtr'] = data['Query'].apply(logical)
```

```python
# Tổng số lần xuất hiện của các từ khóa gây hại trong câu truy vấn
def harmful_keyword_count(x):
    return sum([x.upper().count(kw) for kw in harmful_keywords])
data['harmful_keywords'] = data['Query'].apply(harmful_keyword_count)

# Tổng số lần xuất hiện của các ký tự đặc biệt trong truy vấn
def special_characters(x):
    return len(re.findall(r'[^\w\s]', x))
data['no_spl_chrtr'] = data['Query'].apply(special_characters)
data.head(50)
```

Out[61]:

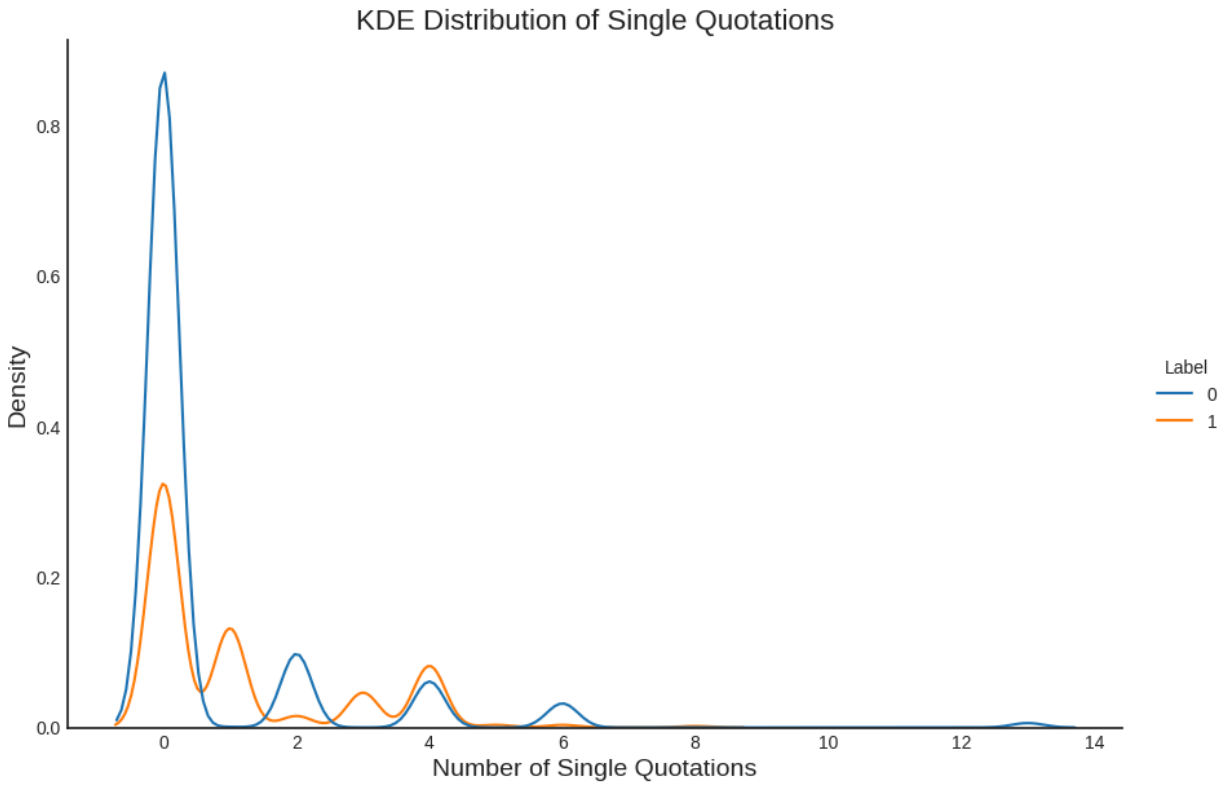
	Query	Label	no_sngle_quts	no_dble_quts	no_punctn	no_sgle_cmnt	no_mlt_c
0	" or pg_sleep (__TIME__) --	1	0	1	10	1	
1	create user name identified by pass123 tempora...	1	0	0	1	0	
2	AND 1 = utl_inaddr.get_host_address (...	1	3	0	25	0	
3	select * from users where id = '1' or @ @1 ...	1	3	0	13	1	
4	select * from users where id = 1 or 1#" (...	1	0	1	10	1	
5	select name from syscolumns where id = ...	1	1	0	7	1	
6	select * from users where id = 1 +$+ or 1 =...	1	0	0	8	1	
7	1; (load_file (char (47,101,116,99,47...	1	0	0	22	0	
8	select * from users where id = '1' or ||/1 ...	1	3	0	14	1	
9	select * from users where id = '1' or \.<\ ...	1	3	0	12	1	
10	? or 1 = 1 --	1	0	0	4	1	
11	) or ('a' = 'a	1	3	0	6	0	
12	admin' or 1 = 1#	1	1	0	3	0	
13	select * from users where id = 1 or " ( ]...	1	0	2	9	1	
14	or 1 = 1 --	1	0	0	3	1	
15	AND 1 = utl_inaddr.get_host_address (...	1	3	0	25	0	
16	select * from users where id = '1' %!<@ uni...	1	3	0	14	1	
17	select * from users where id = 1 or "& (...	1	0	2	9	1	
18	select * from users where id = 1 or "? (...	1	0	2	9	1	
19	distinct	1	0	0	0	0	
20	select * from users where id = '1' * (\ ...	1	3	0	11	1	
21	1 and ascii (lower (substring ((...	1	2	0	14	0	
22	select * from users where id = 1 or \.<\ or ...	1	0	0	7	1	

	Query	Label	no_sngle_quts	no_dble_quts	no_punctn	no_sgle_cmnt	no_mlt_c
23	admin" or "1" = "1"--	1	0	5	8	1	
24	select * from users where id = 1 or "%{" or...	1	0	2	9	1	
25	insert	1	0	0	0	0	
26	select * from users where id = 1 or 1#"? =...	1	0	1	9	1	
27	select * from users where id = 1 or "%?" or...	1	0	2	9	1	
28	AND 1 = utl_inaddr.get_host_address (...	1	3	0	23	0	
29	select * from users where id = 1 or "?#" or...	1	0	2	9	1	
30	or 1 = 1 or "" =	1	0	2	4	0	
31	/**/or/**/1/**/ = /**/1	1	0	0	17	0	
32	select * from users where id = 1 or ". (...	1	0	2	9	1	
33	select * from users where id = '1' union se...	1	3	0	13	1	
34	select * from users where id = '1' union se...	1	3	0	12	1	
35	or 1 = 1/*	1	0	0	3	0	
36	%27 or 1 = 1	1	0	0	2	0	
37	select * from users where id = 1 <@&@ union...	1	0	0	11	1	
38	select * from users where id = '1' union se...	1	3	0	13	1	
39	1; (load_file (char (47,101,116,99,47...	1	0	0	16	0	
40	select * from users where id = '1' or $ 1 ...	1	3	0	12	1	
41	select * from users where id = 1 or \.<1 or ...	1	0	0	7	1	
42	select * from users where id = '1' or \<\ u...	1	3	0	11	1	
43	select * from users where id = 1 union sele...	1	0	0	9	1	
44	select * from users where id = 1 ||1 union ...	1	0	0	9	1	
45	select * from users where id = '1' or \.<\ o...	1	3	0	10	1	

	Query	Label	no_sngle_quts	no_dble_quts	no_punctn	no_sgle_cmnt	no_mlt_c
46	select * from users where id = 1 or "]{" or...	1	0	2	9	1	
47	select * from users where id = 1 or ".@" or...	1	0	2	9	1	
48	select * from users where id = 1 or "$[" or...	1	0	2	9	1	
49	or 1 --'	1	1	0	3	1	

In [62]:
```
plt.figure(figsize=(10, 6))

sns.displot(data=data, x="no_sngle_quts", hue="Label", kind="kde", height=6, aspect=1.

plt.title('KDE Distribution of Single Quotations', fontsize=16)
plt.xlabel('Number of Single Quotations', fontsize=14)
plt.ylabel('Density', fontsize=14)

plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```

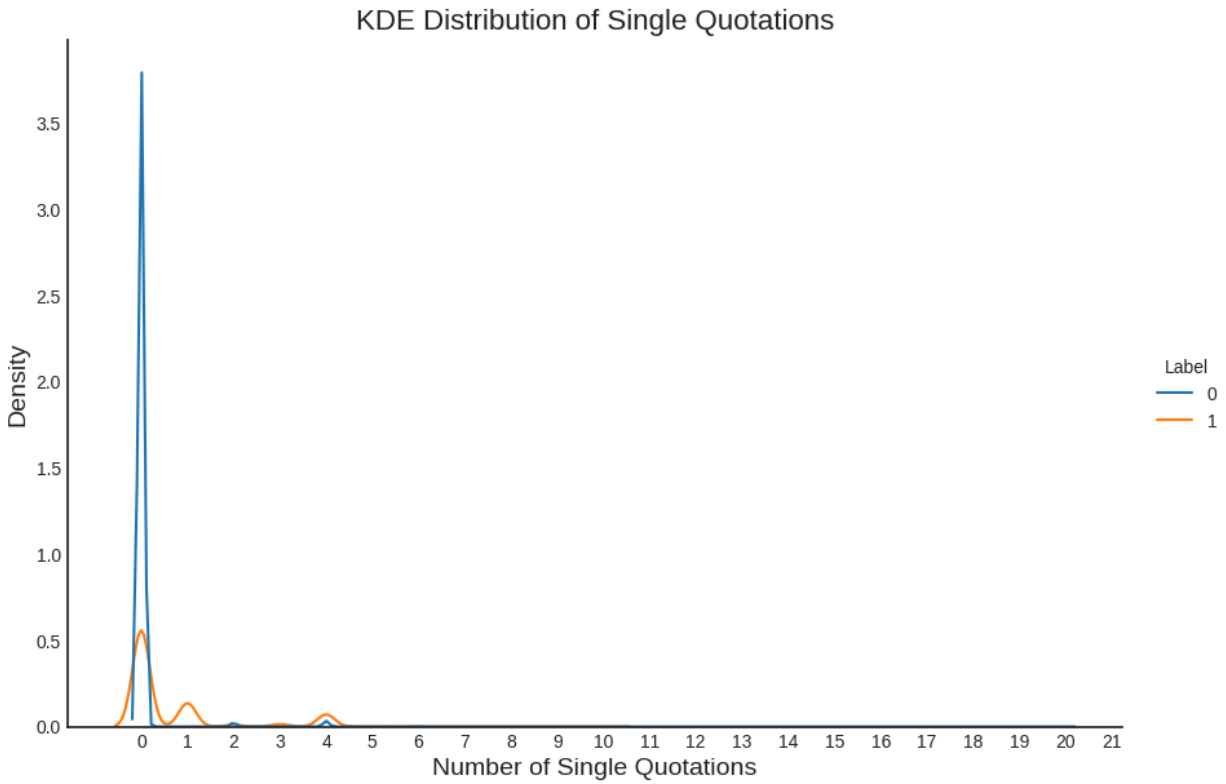


Nhận xét: Với những truy vấn chứa 0 hoặc 2 dấu nháy đơn thì có khả năng truy vấn đó là truy vấn hợp lệ. Ngược lại với những truy vấn chứa 1 hoặc 3 dấu nháy đơn thì có khả năng truy vấn đó là truy vấn có thể bị tấn công.

In [63]:
```
plt.figure(figsize=(10, 6))

sns.displot(data=data, x="no_dble_quts", hue="Label", kind="kde", height=6, aspect=1.5
plt.xticks(range(0, int(data['no_dble_quts'].max()) + 2, 1))
```

```
plt.title('KDE Distribution of Single Quotations', fontsize=16)
plt.xlabel('Number of Single Quotations', fontsize=14)
plt.ylabel('Density', fontsize=14)

plt.show()
```

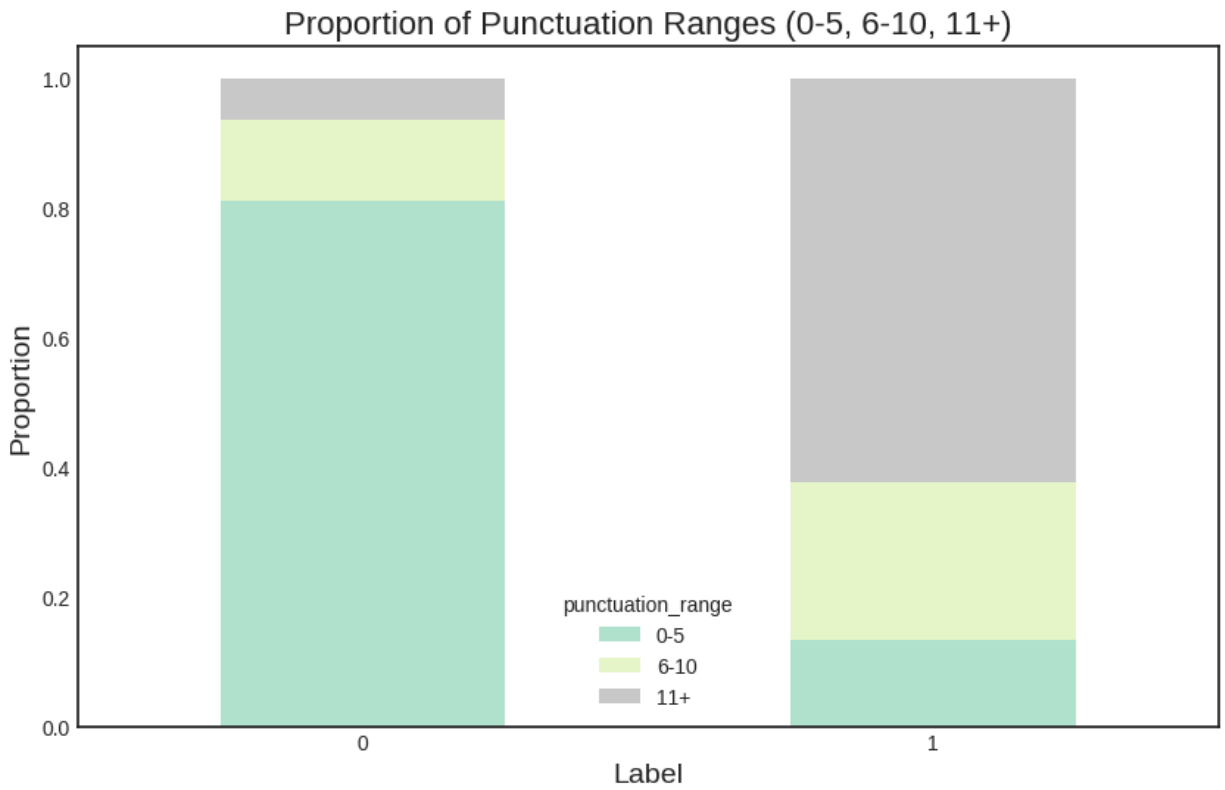
```
<Figure size 1000x600 with 0 Axes>
```



Nhận xét: Với những truy vấn không chứa dấu nháy đơn thì có khả năng đó là những truy vấn hợp lệ, ngược lại với những truy vấn chứa 1 dấu nháy đơn thì có khả năng chúng là những truy vấn độc hại.

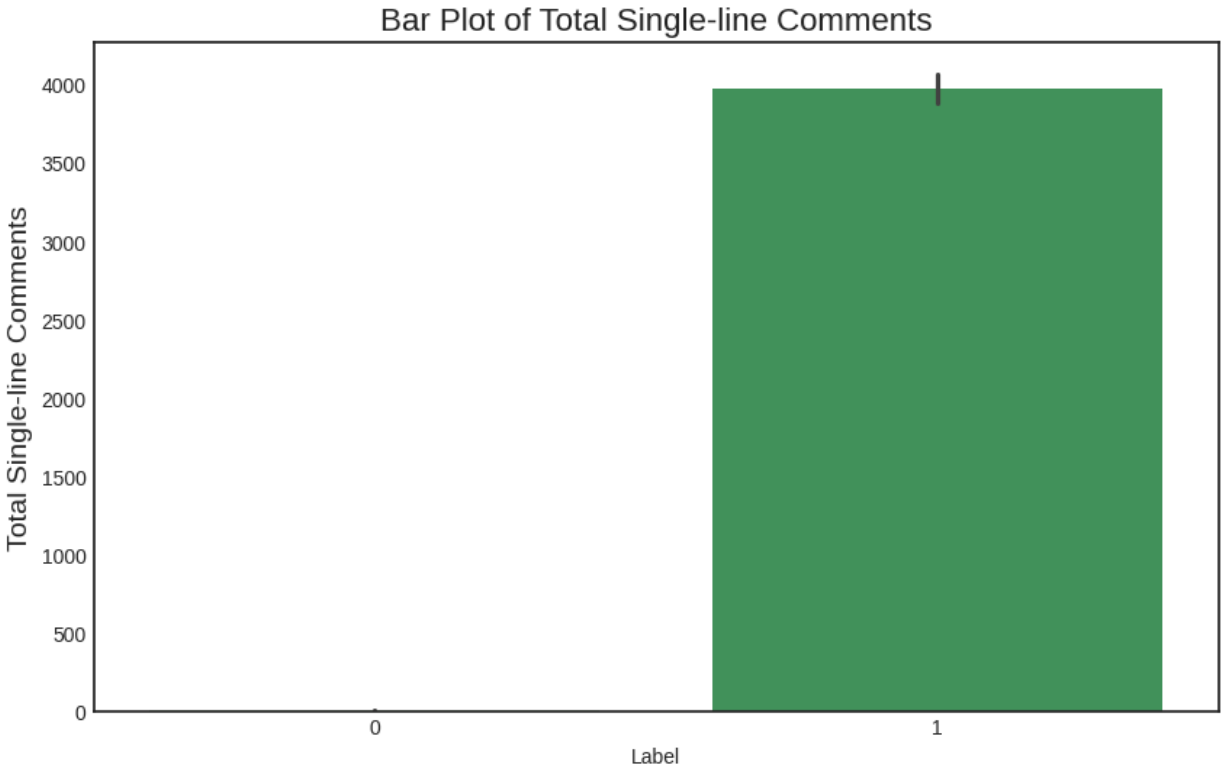
In [64]:
```
bins = [0, 5, 10, data['no_punctn'].max()]
labels = ['0-5', '6-10', '11+']
data['punctuation_range'] = pd.cut(data['no_punctn'], bins=bins, labels=labels, right=

plt.figure(figsize=(10, 6))

punctuation_counts = data.groupby(['Label', 'punctuation_range']).size().unstack(fill_
punctuation_counts.div(punctuation_counts.sum(1), axis=0).plot(kind='bar', stacked=Tru

plt.title('Proportion of Punctuation Ranges (0-5, 6-10, 11+)', fontsize=16)
plt.xlabel('Label', fontsize=14)
plt.ylabel('Proportion', fontsize=14)
plt.xticks(rotation=0)

plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



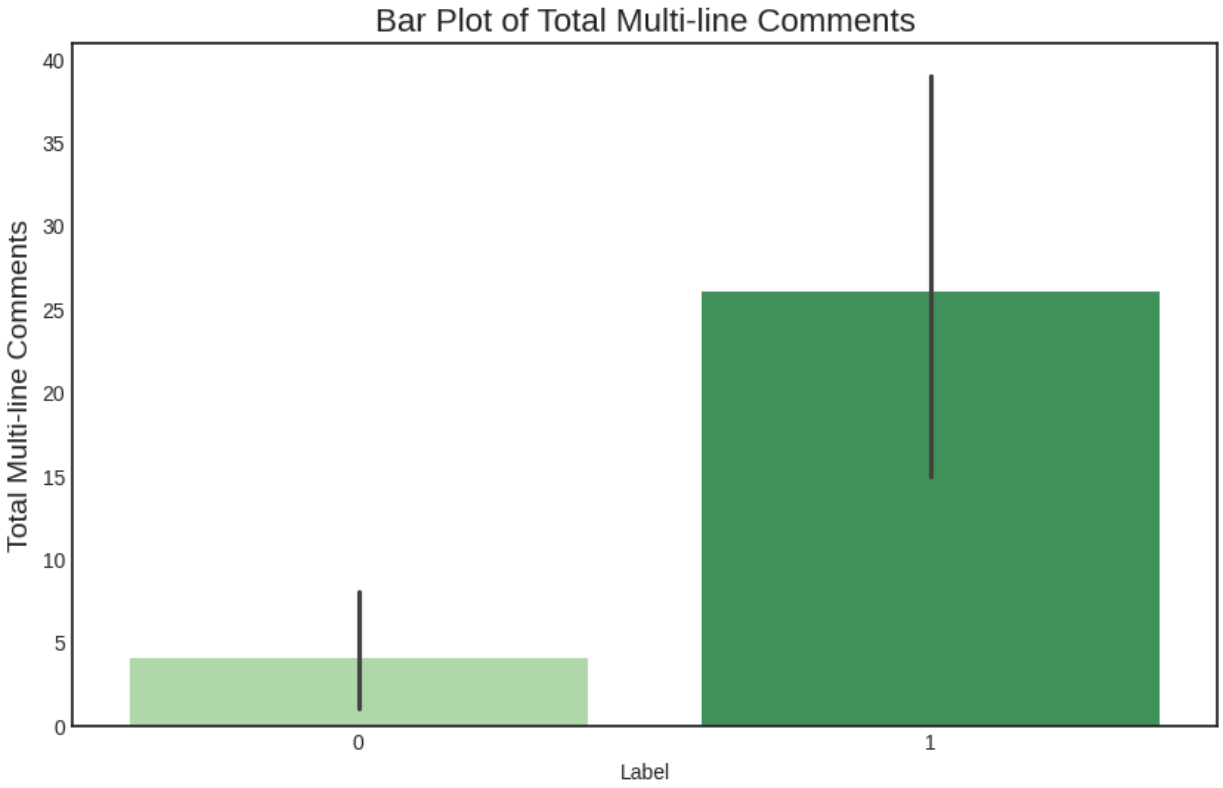
Nhận xét: Số dấu câu tăng lên đồng nghĩa với việc truy vấn có khả năng là injection query.

In [65]:
```
plt.figure(figsize=(10, 6))
sns.barplot(data=data, x="Label", y="no_sgle_cmnt", estimator=sum,palette="Greens")
plt.title('Bar Plot of Total Single-line Comments', fontsize=16)
plt.ylabel('Total Single-line Comments', fontsize=14)
plt.show()
```



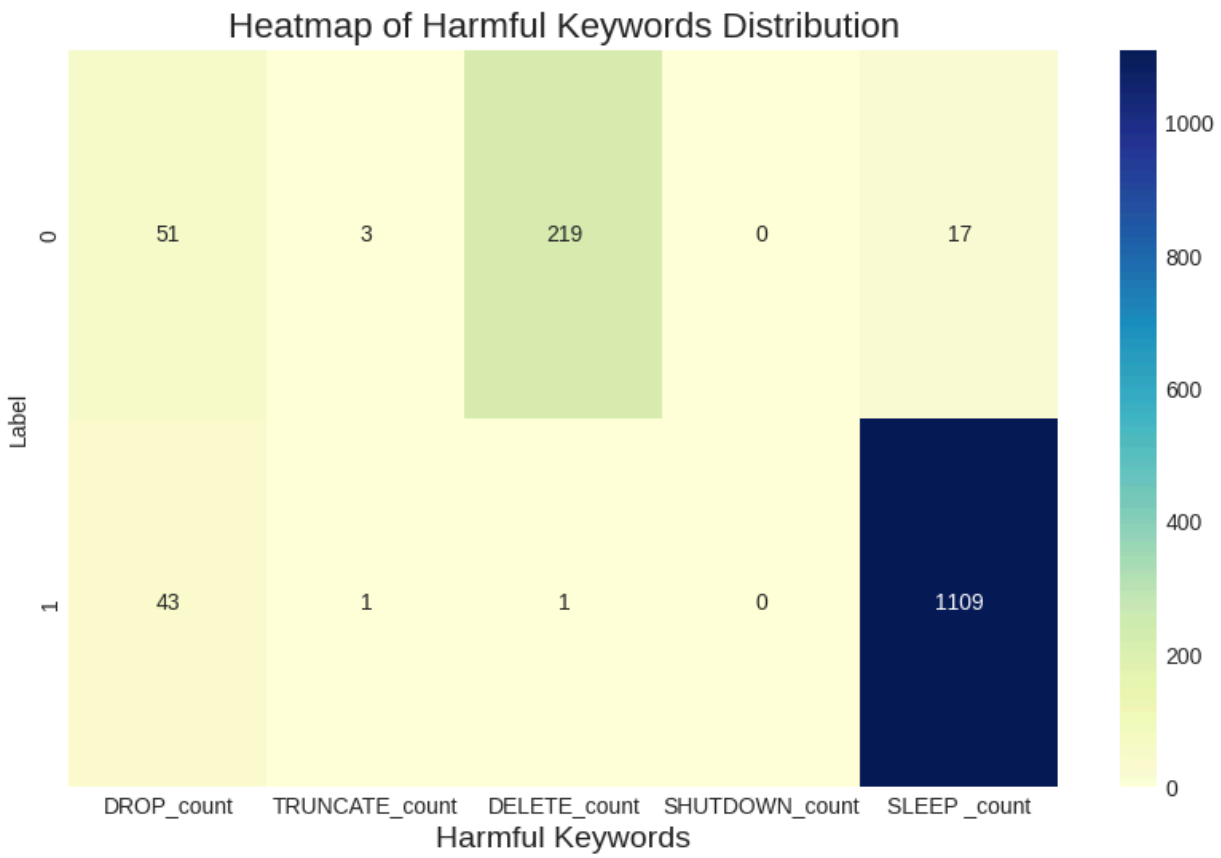
Nhận xét: Truy vấn có single-line comment đều là SQL Injection.

In [66]:
```
plt.figure(figsize=(10, 6))
sns.barplot(data=data, x="Label", y="no_mlt_cmnt", estimator=sum,palette="Greens")
plt.title('Bar Plot of Total Multi-line Comments', fontsize=16)
plt.ylabel('Total Multi-line Comments', fontsize=14)
plt.show()
```



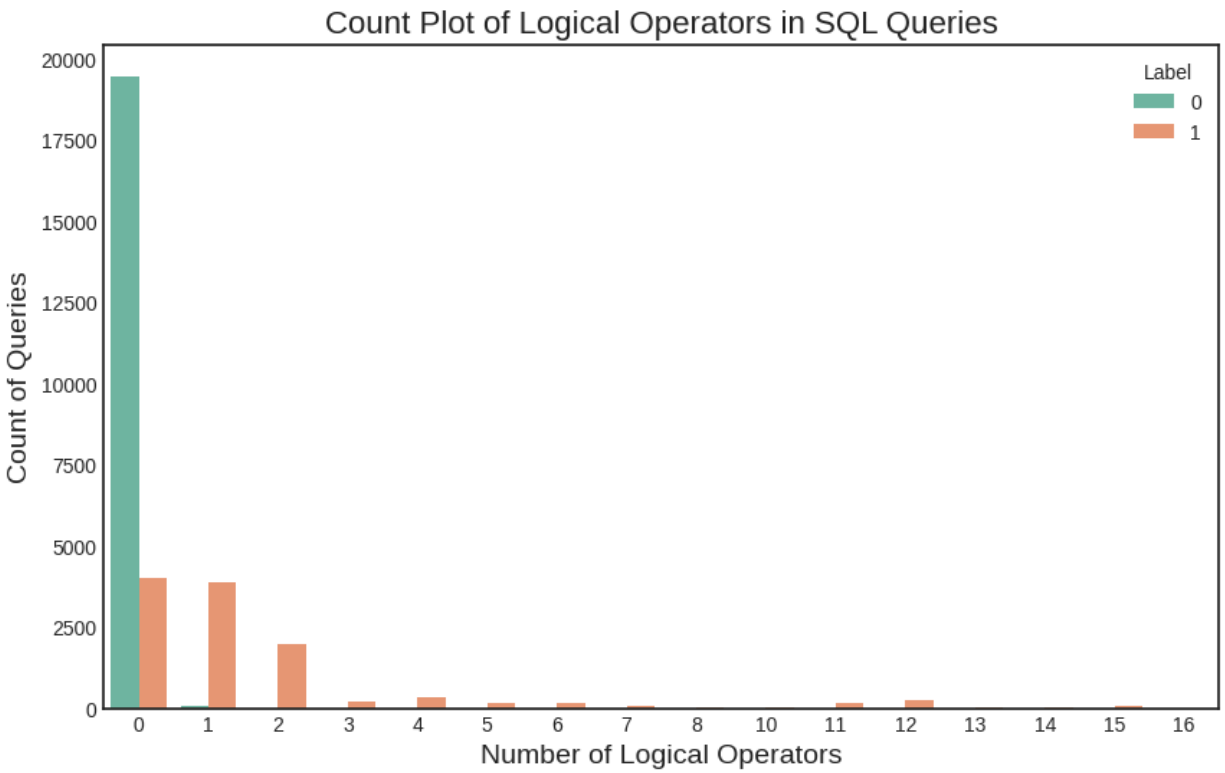
Nhận xét:Multi-line comment ít xuất hiện trong các truy vấn thông thường và thường có xu hướng xuất hiện ở trong các truy vấn SQL injection

In [67]:
```
for kw in harmful_keywords:
    data[f'{kw}_count'] = data['Query'].apply(lambda query: query.upper().count(kw))

harmful_keyword_columns = [f'{kw}_count' for kw in harmful_keywords]
harmful_keywords_data = data.groupby('Label')[harmful_keyword_columns].sum()

plt.figure(figsize=(10, 6))
sns.heatmap(harmful_keywords_data, annot=True, cmap="YlGnBu", fmt='g')

plt.title('Heatmap of Harmful Keywords Distribution', fontsize=16)
plt.xlabel('Harmful Keywords', fontsize=14)

plt.show()
```



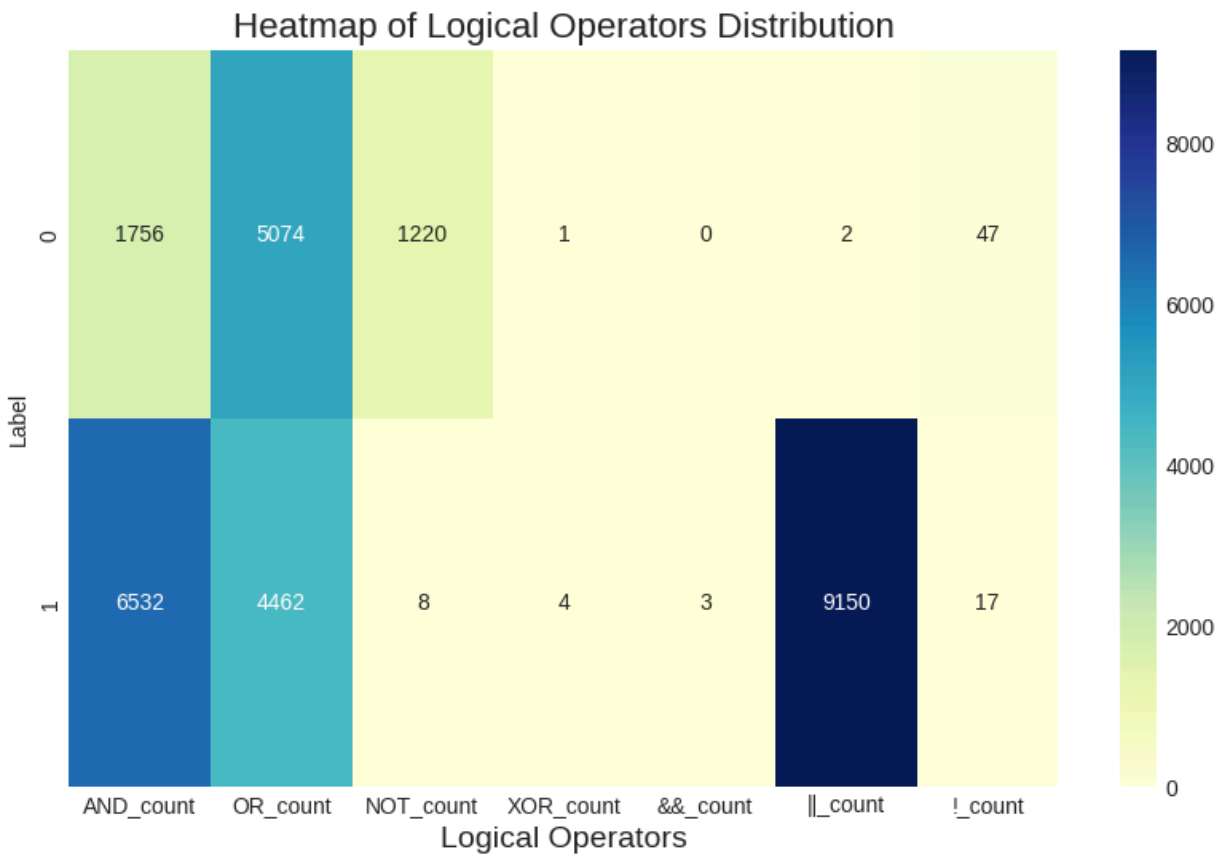
Nhận xét: Các cuộc tấn công SQL Injection sử dụng từ khóa SLEEP rất phổ biến, cho thấy kiểu tấn công time-based chiếm ưu thế, trong khi các từ khóa như DELETE và DROP có thể xuất hiện cả trong câu truy vấn thông thường và SQL Injection.

In [68]:
```
plt.figure(figsize=(10, 6))

sns.countplot(data=data, x="no_log_oprtr", hue="Label", palette="Set2")

plt.title('Count Plot of Logical Operators in SQL Queries', fontsize=16)
plt.xlabel('Number of Logical Operators', fontsize=14)
plt.ylabel('Count of Queries', fontsize=14)

plt.show()
```



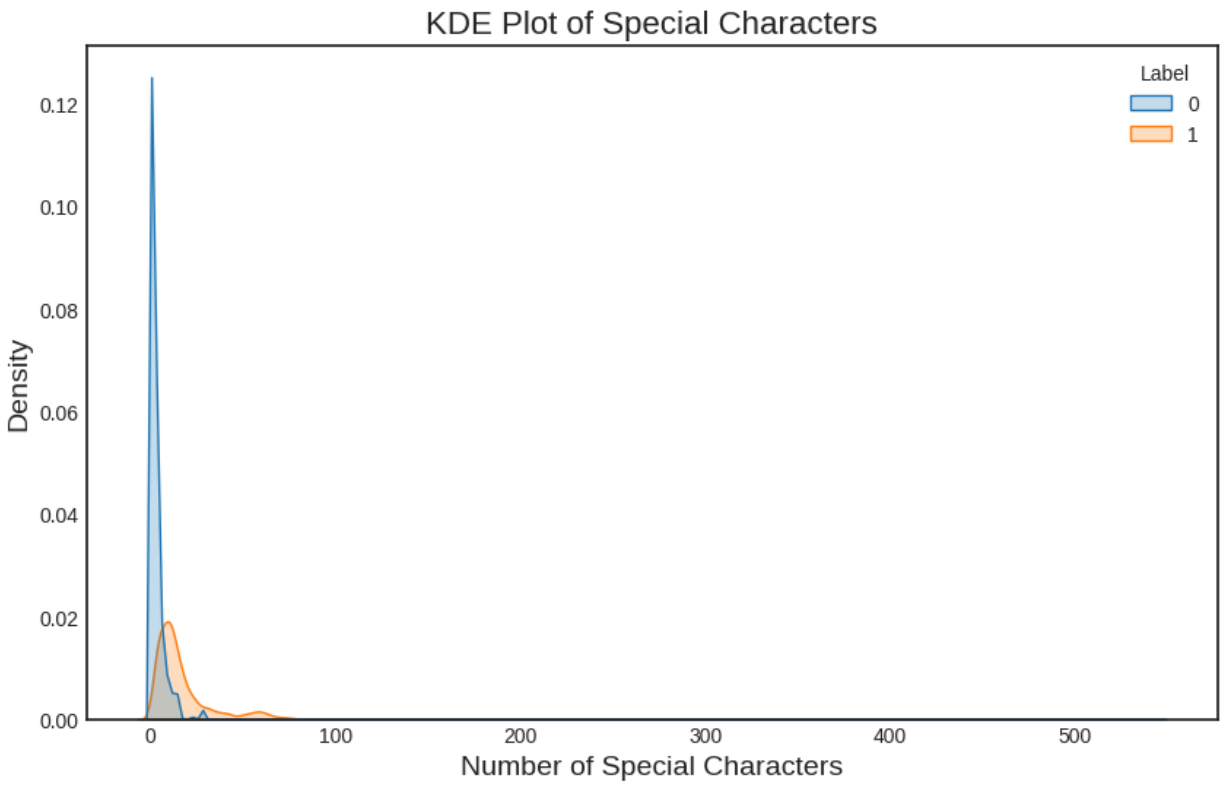
Nhận xét: Các truy vấn có xu hướng sử dụng rất ít hoặc không sử dụng các toán tử logic. Các truy vấn Injection thường sử dụng từ 1 đến 2 toán tử logic để tạo ra các điều kiện phức tạp hơn nhằm khai thác cơ sở dữ liệu. Đáng chú ý, không có nhiều truy vấn với số lượng lớn hơn 2 toán tử logic, điều này cho thấy các tấn công SQL Injection thường giới hạn số lượng toán tử được sử dụng để tránh bị phát hiện.

In [69]:
```
logical_operators = ['AND', 'OR', 'NOT', 'XOR', '&&', '||', '!']
for op in logical_operators:
    data[f'{op}_count'] = data['Query'].apply(lambda query: query.upper().count(op))

logical_op_columns = [f'{op}_count' for op in logical_operators]
logical_op_data = data.groupby('Label')[logical_op_columns].sum()

plt.figure(figsize=(10, 6))
sns.heatmap(logical_op_data, annot=True, cmap="YlGnBu", fmt='g')

plt.title('Heatmap of Logical Operators Distribution', fontsize=16)
plt.xlabel('Logical Operators', fontsize=14)

plt.show()
```



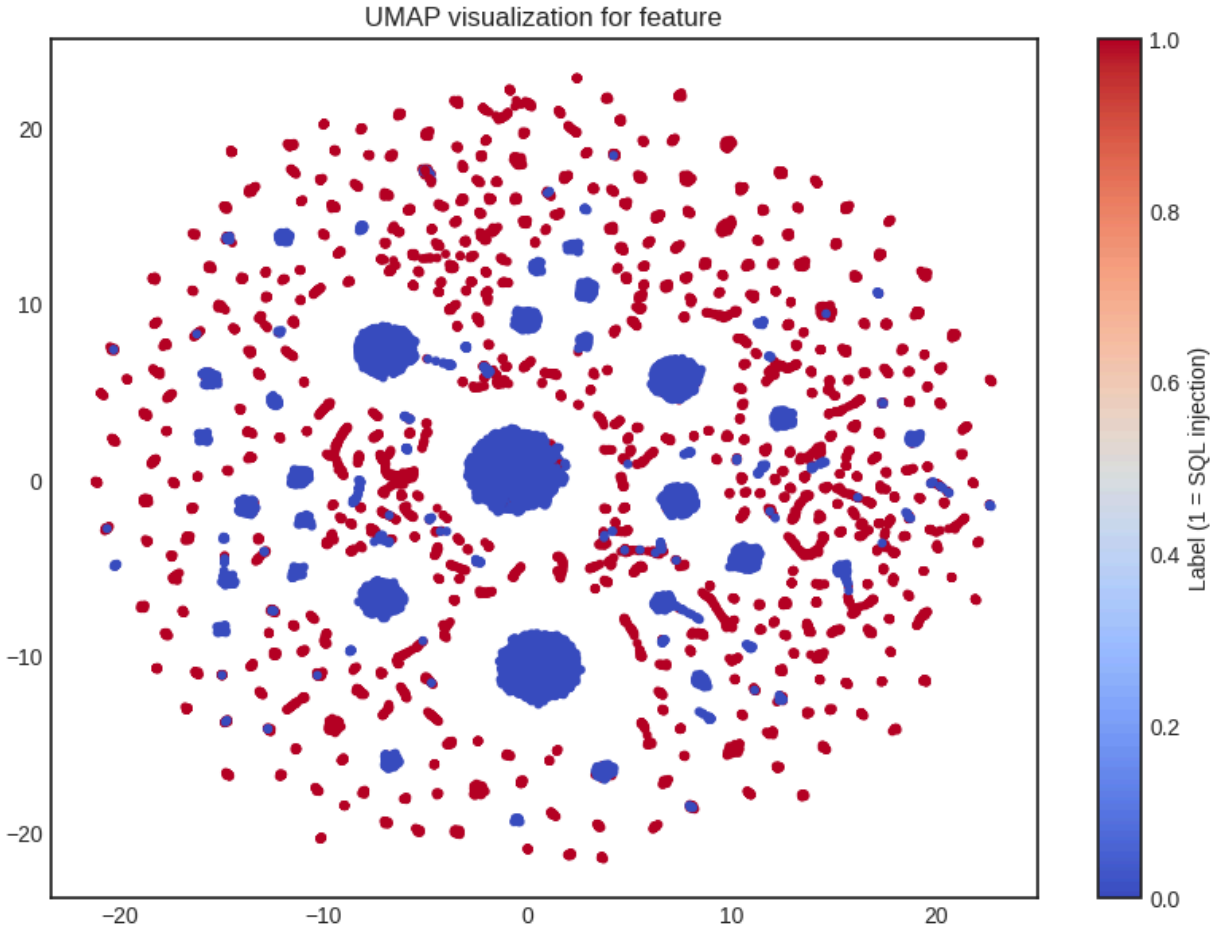
Nhận xét: Các toán tử AND, || được sử dụng với tần suất rất cao, điều này cho thấy kẻ tấn công thường sử dụng các toán tử này để tạo ra các điều kiện phức tạp nhằm khai thác cơ sở dữ liệu.

In [70]:
```
plt.figure(figsize=(10, 6))

sns.kdeplot(data=data, x="no_spl_chrtr", hue="Label", fill=True)

plt.title('KDE Plot of Special Characters', fontsize=16)
plt.xlabel('Number of Special Characters', fontsize=14)
plt.ylabel('Density', fontsize=14)

plt.show()
```



Nhận xét: Các truy vấn SQL Injection có mật độ cao hơn ở vùng có số lượng ký tự đặc biệt lớn hơn(20 đến hơn 100 ký tự đặc biệt). Điều này cho thấy rằng các truy vấn tấn công SQL Injection thường chứa nhiều ký tự đặc biệt hơn. Ngược lại, ở những truy vấn không có ký tự đặc biệt thì hầu hết đều là truy vấn an toàn.

In [72]:
```
from sklearn.preprocessing import StandardScaler
import umap
import matplotlib.pyplot as plt


features = data[['no_sngle_quts', 'no_dble_quts', 'no_punctn', 'no_sgle_cmnt',
                 'no_mlt_cmnt', 'no_log_oprtr', 'harmful_keywords', 'no_spl_chrtr']]

# Normalization
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

# Apply UMAP
umap_model = umap.UMAP(n_components=2, n_neighbors=10, min_dist=0.5, random_state=42)
umap_embedding = umap_model.fit_transform(scaled_features)

# Visualization
plt.figure(figsize=(10, 7))
plt.scatter(umap_embedding[:, 0], umap_embedding[:, 1], c=data['Label'], cmap='coolwar
plt.title("UMAP visualization for feature")
plt.colorbar(label='Label (1 = SQL injection)')
plt.show()
```



Nhận xét: Mặc dù đồ thị có các điểm chồng lắp nhưng hầu hết các nhãn đều phân biệt.

In [73]:
```
import re
import nltk
nltk.download('punkt')
nltk.download('wordnet')
import string
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve,auc
from nltk.stem.porter import PorterStemmer
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
from gensim.models import word2vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

In [74]:
```
#printing some random variable
sent_0 = data['Query'].values[0]
print(sent_0)
print("="*50)

sent_2000 = data['Query'].values[2000]
print(sent_2000)
print("="*50)

sent_15000 = data['Query'].values[15000]
print(sent_15000)
print("="*50)

sent_20000 = data['Query'].values[20000]
print(sent_20000)
print("="*50)
```

```
" or pg_sleep  (  __TIME__  )  --
==================================================
1   )    )     )   or exp  (  ~  (  select * from   (  select concat  (  0x7171706a7
1,  (  select   (  elt  (  6270  =  6270,1   )    )     )  ,0x717a767a71,0x78   )
)   x   )    )    and    (    (     (  2230  =  2230
==================================================
6173
==================================================
  SELECT SUBSTRING_INDEX ( "www.w3schools.com", ".", 1 ) ;
==================================================
```

In [75]:
```
#remove the special Character :  https://stackoverflow.com/a/5843547/4084039
sent_15000 = re.sub('[^A-Za-z0-9]+',' ',sent_15000)
print(sent_15000)
```

```
6173
```

In [76]:
```
sent_20000 = re.sub('[^A-Za-z0-9]+',' ',sent_20000)
print(sent_20000)
```

```
 SELECT SUBSTRING INDEX www w3schools com 1
```

Stop words are common words that will likely appear in any text. Which we need to remove.

E.g.: silver or lead is fine for me -> silver, lead, fine.

In [77]:
```
stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselve
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him',
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', '
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "t
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'h
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'o
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'ar
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'n
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't"
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mig
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", '
            'won', "won't", 'wouldn', "wouldn't"])
```

In [78]:
```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_query = []
lemmatizer = WordNetLemmatizer()
# tqdm is for printing the status bar
for sentance in tqdm(data['Query'].values):
    sentance = re.sub('[^A-Za-z0-9]+', ' ', sentance)
    sentance = re.sub(r',', ' ', sentance)
    #https://www.machinelearningplus.com/nlp/lemmatization-examples-python/
    tokenization = nltk.word_tokenize(sentance)
    sentance = ' '.join([lemmatizer.lemmatize(w) for w in tokenization])
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not in stopwo
    preprocessed_query.append(sentance.strip())
```

```
100%|██████████| 30903/30903 [00:07<00:00, 4231.30it/s]
```

In [79]:
```
preprocessed_query
```

Out[79]:
```
['pg sleep time',
 'create user name identified pass123 temporary tablespace temp default tablespace us
er',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 5',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 'select name syscolumns id select id sysobjects name tablename',
 'select user id 1 1 1 1',
 '1 load file char 47 101 116 99 47 112 97 115 115 119 100 1 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 '1 1',
 '',
 'admin 1 1',
 'select user id 1 1 1 1',
 '1 1',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 5',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'distinct',
 'select user id 1 1 1 1',
 '1 ascii lower substring select top 1 name sysobjects xtype u 1 1 116',
 'select user id 1 1 1 1',
 'admin 1 1',
 'select user id 1 1 1 1',
 'insert',
 'select user id 1 1 1 1 1 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 8',
 'select user id 1 1 1 1',
 '1 1',
 '1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select version 1',
 '1 1',
 '27 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select version 1',
 '1 load file char 47 101 116 99 47 112 97 115',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select 1 1 version 1',
 'select user id 1 1 union select null banner v version rownum 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 '1',
 'select user id 1 1 1 1',
 'admin 1 1',
 'select user id 1 1 union select null banner v version rownum 1 1',
 'x27union select',
 'print',
 'select user id 1 union select version 1',
```

```
 '1 1',
 '1 pg sleep time',
 'select user id 1 1 1 1',
 'admin 1 1',
 '2 1 3',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select null version 1',
 'admin 1 1',
 'not select system user sa waitfor delay 0 0 2',
 'select user id 1 union select 1 banner v version rownum 1 1',
 '0 0',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 'char 4039 41 2b 40select',
 'select user id 1 1 1 1 1',
 'admin 1 1',
 'declare varchar 200 select 0x73656c6',
 '0 0',
 'select user id 1 1 1 1',
 'x x',
 'select user id 1 union select 1 version 1',
 'sleep 50',
 'select user id 1 1 union select null banner v version rownum 1 1',
 '1 union select 1 2 3 4 5 6 name sysobjects xtype u',
 'select user id 1 1 union select version 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 1 1 1 1',
 '1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 4',
 'select user id 1 union select version 1',
 '1 user name dbo',
 'not substring select version 24 1 1 waitfor delay 0 0 2',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 6',
 '1 1',
 'select user id 1 1 1 1',
 'text',
 '1234 1 0 union select admin 81dc9bdb52d04dc20036dbd8313ed055',
 'select user id 1 1 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'admin 1 1',
 'select user id 1 1 1 1 1',
 '',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 4',
 'select information schema table',
 'select user id 1 1 1 1',
 '1 pg sleep time',
 'select user id 1 union select version 1',
 'declare q nvarchar 200 select q 0x770061',
```

```
 '1 1',
 'select user id 1 1 1 1 1',
 'declare q nvarchar 200 select q 0x770061006900740066006f00720020006400650065006c0061007
900200027003000 3a0030003a0031003000270000 exec q',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'unusual unusual',
 '1 select version',
 'truncate',
 '',
 '3 3',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'select user id 1 union select 1 version 1',
 'admin 1 1',
 'select user id 1 1 union select version 1',
 '1 benchmark 10000000 md5 1',
 'select user id 1 union select 1 version 1',
 '1 utl inaddr get host address select sys database name dual',
 '0 0',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 'not substring select version 25 1 0 waitfor delay 0 0 2',
 '1 sleep time',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 'true',
 'execute immediate sel ect u er',
 'hi x x',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'admin',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'union select 1 load file etc passwd 1 1 1',
 'waitfor delay 0 0 time',
 'unusual unusual',
 'insert mysql user user host password value name localhost password pass123',
 'admin 1 1',
 'select user id 1 1 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'like',
 'username like char 37',
 'pg sleep time',
 'select user id 1 1 1 1',
 '3 3',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 'exec sp',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'admin 1 1',
 'waitfor delay 0 0 10',
 'select user id 1 union select null version 1',
 'select user id 1 union select 1 version 1',
 '1 select var temp',
 'hi',
```

```
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 1',
 'admin 1 1',
 'true',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '1 pg sleep time',
 'select user id 1 1 1 1',
 'sleep time',
 'select user id 1 1 1 1',
 'admin 1 1',
 'select user id 1 union select 1 version 1',
 'exec xp regread',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 '1 1 1',
 '0 0',
 'union select user login char',
 'select user id 1 union select 1 version 1',
 'admin 1 1',
 '1 select',
 'select user id 1 1 1 1',
 'x 1 select count tabname',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 '1 benchmark 10000000 md5 1',
 'procedure',
 '',
 '1 utl inaddr get host address select sys login user dual',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 2',
 'desc user',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1 1 1',
 '',
 '',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'admin',
 'select user id 1 1 1 1',
 'utl http request',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 1',
 'select user id 1 union select 1 version 1',
 'benchmark 10000000 md5 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '1 uni select',
 'select user id 1 union select null version 1',
 'select user id 1 1 union select null version 1',
 'admin 1 1',
 'utl http request http 192 168 1 1',
 '7659 7659',
```

```
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 6',
 'select user id 1 1 1 1',
 'true',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select version 1',
 'select user id 1 1 union select 1 version 1',
 'true',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '3 3',
 '1 utl inaddr get host address select global name global name',
 'select user id 1 1 union select null banner v version rownum 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 1 union select version version 1',
 'select user id 1 1 1 1 1',
 'uid like',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 '23 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select version 1',
 '0 1 1',
 'select user id 1 1 1 1',
 'admin 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 5',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 7',
 'apos',
 'sqlvuln',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 '',
 'hi 1 1',
 'select user id 1 union select null banner v version rownum 1 1',
 'exec master xp cmdshell ping 172 10 1 255',
 'select user id 1 1 1 union select 1 banner v version rownum 1 1',
 'select user id 1 union select version 1',
 'select information schema table',
 'select user id 1 1 1 1',
 'select top 1',
 '1 1',
 '',
 '1 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 union select version 1',
 '',
 'exec sp addlogin name password',
 'select user id 1 union select null banner v version rownum 1 1',
 '0x77616974666f722064656c61792027303a303a31302700 exec',
```

```
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'group userid 1 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1 1',
 '1 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 '',
 'declare varchar 22 select',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 union select version version 1',
 'variable',
 'exec master xp cmdshell',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select 1a version 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 '1 select var temp',
 'select user id 1 union select 1 banner v version rownum 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 'x x',
 '',
 '1 1',
 'select user id 11 1 union select 1 version 1',
 'x x',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 4',
 'srvrolemember sysadmin 0 waitfor delay 0 0 2',
 'select user id 1 1 1 1',
 'select user id 1 union select version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select version version 1',
 '1 1',
 '',
 'x x',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'password',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select null version 1',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 3',
 '1 1',
 'admin 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 'union select user login char 114 111 111 116',
 'select user id 1 1 union select 1 version 1',
 '1 1',
```

```
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 5',
 'pg sleep time',
 '3 3',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'union select information schema table',
 'select user id 1 1 1 1',
 'benchmark 10000000 md5 1',
 '1 utl inaddr get host address select count distinct column name sys tab column',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 7',
 '',
 'delete',
 'waitfor delay 0 0 time',
 '1 1',
 '1 non existant table 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'declare varchar 200 select 0x77616974',
 'x userid null',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select version 1',
 'select user id 1 union select 1 version 1',
 '1 utl inaddr get host address select host name v instance',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 2',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'union select',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'x full name like bob',
 'x member email null',
 '2 1',
 'select user id 1 1 1 1 1',
 'pg sleep time',
 '',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'sleep time',
 'sleep 50',
 'select user id 1 1 1 union select 1 banner v version rownum 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'sleep time',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select version 1',
 'select user id 1 union select version 1',
 'select user id 1 union select null version 1',
 '2 1',
 'exec sel ect u er',
 'anything x x',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select 1 banner v version rownum 1 1',
```

```
 'select user id 1 union select null version 1',
 '',
 '3 3',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'admin 1 1',
 'uni sel ect',
 '',
 'waitfor delay 0 0 time',
 'select user id 1 1 1 1',
 '1 waitfor delay 0 0 10',
 'bfilename',
 'admin 1 1',
 'uef',
 'password',
 '1 sleep time',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 6',
 'select user id 1 union select version 1',
 'select user id 1 1 union select 1 version 1',
 '1 1 select count tablenames',
 'sleep time',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 8',
 'select user id 1 1 1 1 1 1',
 '6',
 '1',
 '1 load file char 110 46 101 120 116 char 39 39 1 0',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '1 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 7',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'whatever whatever',
 'declare varchar 200 select 0x73656c65637420404076657273696f6e exec',
 'select user id 1 1 1 1',
 '',
 '1 utl inaddr get host address select count distinct table name sys table',
 'select user id 1 union select 1 version 1',
 'admin 1 1',
 'select user id 1 1 union select version 1',
 'select user id 1 1 1 1',
 '',
 'select user id 1 1 1 1',
 '1 1',
 'drop table temp',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 2',
 'benchmark 10000000 md5 1',
 'x x',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'something like',
 'not select serverproperty isintegratedsecurityonly 0 waitfor delay 0 0 2',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 4',
```

```
 'select user id 1 1 1 1',
 'exec master xp cmdshell ping 10 10 1 2',
 'admin 1 1',
 'not select serverproperty isintegratedsecurityonly 1 waitfor delay 0 0 2',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 2',
 'select user id 1 union select version version 1',
 'select user id 1 1 union select null version 1',
 'waitfor delay 0 0 time',
 'x userid null',
 'select user id 1 1 1 1 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 '1 1',
 'select user id 1 union select 1a banner v version rownum 1 1',
 'admin 1 1',
 'select user id 1 1 union select 1 version 1',
 '1 load file char 110 46 101 120 11',
 'exec',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 '1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 union select version version 1',
 'x email null',
 'select user id 1 1 union select 1 version 1',
 '',
 'select user id 1 union select null banner v version rownum 1 1',
 'select user id 11 1 union select 1 banner v version rownum 1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'update',
 '1',
 'select user id 1 1 union select 1 version 1',
 'pg sleep time',
 '6',
 'select user id 1 union select 1 version 1',
 'exists',
 '',
 'text n text',
 'select user id 1 union select null version 1',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 '0 0',
 'select user id 1 1 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select version version 1',
 'select user id 1 1 union select version 1',
 'password 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'union select null select version',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select null version 1',
```

```
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 '1 utl inaddr get host address select banner v version rownum 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 3',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'union select',
 'asc',
 'myappadmin adduser admin newpass',
 'exec sp addsrvrolemember name sysadmin',
 'select user id 1 1 union select 1 version 1',
 'true',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 3',
 'userid like',
 'union select null null select version',
 '',
 'select user id 1 1 1 1',
 'admin',
 'select user id 1 union select 1 version 1',
 'waitfor delay 0 0 time',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 'admin 1 1',
 'union select version',
 '',
 'sleep time',
 'union select null null null null null select version',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 1 1 1',
 'hi',
 'select name syscolumns id sele',
 'select user id 1 1 1 1 1',
 'not substring select version 25 1 5 waitfor delay 0 0 2',
 'select user id 1 union select 1 version 1',
 '2 1 3',
 'union select version',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 8',
 'exec master xp cmdshell nslookup www google com',
 'isnull 1 0',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'declare q nvarchar 200 0x730065006c00650063',
 '',
 'union select version',
 'select user id 1 1 1 union select 1 version 1',
 'admin 1 1',
 'exists',
```

```
 'select user id 1 union select 1 version 1',
 'declare varchar 8000 select 0x73656c',
 'select user id 1 union select version 1',
 'select user id 1 union select 1 banner v version rownum 1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 1a version 1',
 'begin declare var varchar 8000 set var select var var login password user login',
 'union select null null null null select version',
 'hi 1 1',
 'union select',
 'select user id 1 1 1 1',
 'select user id 1 1 union select version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 'desc',
 'select user id 1 1 1 1 1',
 'anything x x',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '0x770061006900740066006f007200200064006500 6c00',
 'select user id 1 union select version version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'isnull 1 0',
 'insert user login password level value char 0x70 char 0x65 char 0x74 char 0x65 char
0x72 char 0x70 char 0x65 char 0x74 char 0x65 char 0x72 char 0x64',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 1',
 'admin 1 1',
 'admin 1 1',
 'user like',
 'admin 1 1',
 'sleep time',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 '',
 'select user id 1 union select version 1',
 'admin 1 1',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 3',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'x 1 1 x',
 '1 utl inaddr get host address select distinct table name select distinct table name
rownum limit sys table limit 8',
 'select user id 1 union select 1 version 1',
 'x x',
 '',
 'select user id 1 1 union select version version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select null version 1',
 'select user id 1 union select version 1',
 'select user id 1 1 union select null version 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 union select null version 1',
 'sleep time',
 '1 1',
 '',
```

```
 'select user id 1 1 1 1 1 1',
 'group userid 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 '0 0',
 '0 0',
 'pg sleep time',
 '1 1',
 'select user id 1 1 1 union select version 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 '1 utl inaddr get host address select count distinct username sys users',
 'sqlvuln',
 'select user id 1 1 1 1',
 'select user id 1 1 union select version 1',
 '0x730065006c0065006300740020004000400076006500',
 'select user id 1 union select 1 1 version 1',
 'select user id 1 union select version version 1',
 'declare varchar 8000 select 0x73656c656374204040766572736f6e',
 'select user id 1 1 1 1',
 '',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct column name select distinct column na
me rownum limit tab column limit 6',
 'select user id 1 union select 1 version 1',
 'username like char 37',
 'select user id 1 1 union select version version 1',
 'something thing',
 '1 select version',
 'declare q nvarchar 4000 select q',
 'select user id 1 union select 1 version 1',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 2',
 'select user id 1 1 1 1',
 'select user id 1 union select null version 1',
 'select user id 1 1 1 1 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 1',
 'select user id 1 1 union select version version 1',
 'select user id 1 union select 1a banner v version rownum 1 1',
 'select user id 1 1 union select 1 version 1',
 'admin 1 1',
 '',
 'select user id 1 1 1 1',
 '1 1',
 'declare q nvarchar 200 0x730065006c0065006300740020004000400076006500720073006900f
006e00 exec q',
 'admin',
 'select user id 1 1 1 1 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'begin declare var varchar 8000 set var',
 'select user id 1 1 1 1',
 'pg sleep time',
 'print variable',
 'select user id 1 1 1 1 1',
 '',
 'x x',
 'select user id 1 1 1 1 1',
 '1 1',
```

```
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '',
 '2a 28 7c 28mail 3d 2a 29 29',
 'select user id 1 1 1 1',
 '1 1',
 'x 1 select count tabname',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 3',
 'select user id 1 union select 1 version 1',
 'admin 1 1',
 '',
 'select user id 1 1 1 union select null version 1',
 'select user id 1 union select 1 banner v version rownum 1 1',
 'x x',
 'union select',
 '23 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'username like',
 'elt 3 5 bin 15 ord 10 hex char 45',
 'uname like',
 'declare varchar 200 select 0x77616974666f722064656c61792027303a303a31302700 exec',
 '1',
 'select user id 1 union select null version 1',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 1',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 1',
 'select user id 1 1 1 1',
 'admin 1 1',
 '1 sleep time',
 'select user id 1 1 1 1',
 'sqlattempt2',
 'select user id 1 union select null version 1',
 'x email null',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 8',
 'select user id 1 1 union select version version 1',
 '1234 1 0 union select admin 81dc9bdb52d04dc20036dbd8313ed055',
 'select user id 1 union select 1 version 1',
 'var select var var temp end',
 'select user id 1 union select null version 1',
 '',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select version 1',
 '',
 'waitfor delay 0 0 time',
 '',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 7',
```

```
 'select user id 1 union select 1 version 1',
 'union select',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'exec master xp cmdshell nslookup www googl',
 'benchmark 10000000 md5 1',
 'union select',
 'select user id 1 1 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '',
 'x member email null',
 'select',
 '1 1',
 '',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 'admin 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 union select 1 banner v version rownum 1 1',
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 5',
 '1 utl inaddr get host address select count distinct granted role dba role privs gra
ntee sys login user',
 'select user id 1 1 union select null version 1',
 '1 utl inaddr get host address select distinct password select distinct password row
num limit sys user limit 6',
 'select user id 1 1 1 1',
 'union select null null null select version',
 'x x',
 'select user id 1 union select version 1',
 '1 1',
 'select user id 1 1 1 1',
 '3 3',
 '1 benchmark 10000000 md5 1',
 '1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'admin 1 1',
 'text',
 '',
 'select user id 1 1 union select version 1',
 'benchmark 10000000 md5 1',
 'exec xp',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1 1',
 'hi',
 '',
 'sqlattempt1',
 'admin 1 1',
 'select user id 1 1 1 1',
 '1 1 1',
 'select user id 1 1 union select 1 banner v version rownum 1 1',
 'variable',
```

```
 '2a 7c',
 'select user id 1 union select 1 version 1',
 'mail',
 'not substring select version 25 1 8 waitfor delay 0 0 2',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 1 1 union select version version 1',
 '1 utl inaddr get host address select count distinct password sys user',
 'select user id 1 1 union select 1 version 1',
 'order',
 'select user id 1 1 1 1',
 'username not null username',
 'admin 1 1',
 '',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select version version 1',
 'select user id 1 1 1 1 1',
 'select user id 1 union select 1 version 1',
 '',
 '1 0 union',
 '',
 '0x770061006900740066006f007200200064006500 6c006100790020002700300003a0030003a',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 union select 1 banner v version rownum 1 1',
 'sqlvuln',
 'text n text',
 '1 1',
 '0x730065006c006500630074002000400040007600650072007300690006f006e00 exec q',
 'select user id 1 1 union select 1 version 1',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1 1 1',
 'select user id 1 1 1 1',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 '1 exec sp exec xp',
 'select user id 1 1 1 union select 1 banner v version rownum 1 1',
 '3 3',
 'admin',
 'select top 1',
 'select user id 1 1 1 1 1',
 '',
 'benchmark 10000000 md5 1',
 'select user id 1 1 1 1',
 '7659 7659',
 'select user id 1 1 union select version version 1',
 'select user id 1 1 1 1 1',
 '1 1 1',
 '',
 'replace',
 'select user id 1 union select 1 version 1',
 'select user id 1 1 1 1',
 'select user id 1 union select null version 1',
 'select user id 1 union select 1a banner v version rownum 1 1',
 '',
 'objectclass',
 '1 utl inaddr get host address select distinct granted role select distinct granted
role rownum limit dba role privs grantee sys loginuser limit 4',
 'limit',
 '1 1',
 'select user id 1 1 1 1',
 '',
```

```
 '1 utl inaddr get host address select distinct username select distinct username row
num limit sys users limit 7',
 'create user name identified pass123',
 'select user id 1 union select null version 1',
 'select user id 1 1 union select version 1',
 '',
 'select user id 1 1 1 1 1',
 '',
 'select user id 1 1 1 1 1',
 'x full name like bob',
 'select user id 1 union select version 1',
 'select user id 1 1 1 1',
 'select user id 1 union select version 1',
 'benchmark 10000000 md5 1',
 'waitfor delay 0 0 time',
 'sleep time',
 'wapiti',
 'sleep time',
 'sleep time',
 'pg sleep time',
 'sleep time',
 'sleep time',
 'pg sleep time',
 'pg sleep time',
 'sleep time',
 'waitfor delay 0 0 time',
 '1 8156 select count generate series 1 5000000',
 '1 clye 7842 7842 char 109 char 79 char 70 char 90 regexp substring repeat right cha
r 5012 0 5000000000 null',
 '4860 azyx 6901 6901 union select 6901 6901 6901 6901 6901',
 '1 union select null null null null null',
 'select benchmark 5000000 md5 0x4c4d6142 6866 6866',
 '1 boed 6787 6787',
 '1 vdbf 7969 7969 extractvalue 1297 concat 0x5c 0x7171706a71 select elt 1297 1297 1
0x717a767a71',
 '1 4281 4281',
 '1 8024 3560',
 '1 5452 6050 6050 ciyc like ciyc',
 '1 8148 like abcdefg upper hex randomblob 500000000 2',
 '1 select case 9443 9443 sleep 5 else 9443 select 9443 information schema character
set end',
 '1 select rawn dual 4988 4988 char 68 char 69 char 97 char 85 regexp substring repea
t right char 5389 0 5000000000 null',
 '4925 union select 5686 5686 5686 5686 5686 5686 5686',
 '1 dnhd 2657 2657 4240 select 4240 pg sleep 5',
 '9534 3038 3038',
 '1 6941 6941 6537 dbms pipe receive message chr 76 chr 116 chr 117 chr 65 5',
 'select case 7978 6009 7978 else 1 select 0 end',
 '1212 make set 7588 2306 2306',
 '1 5466 5466 2388 benchmark 5000000 md5 0x6d457153',
 '3794 union select 2485 2485 2485 2485 2485',
 '1 5286 select count user t1 user t2 user t3 user t4 user t5 gmil gmil',
 '1 8635 select count generate series 1 5000000',
 '6400 union select 4650 4650 4650',
 '1 select rttq dual 7368 7368 updatexml 1808 concat 0x2e 0x7171706a71 select elt 180
8 1808 1 0x717a767a71 8666',
 '1 ztkr 1532 1532',
 '1 9842 9842 union select null null null null null null null',
 '1 exp select select concat 0x7171706a71 select elt 6270 6270 1 0x717a767a71 0x78 x
fbsi like fbsi',
```

```
 '1 char 68 char 69 char 97 char 85 regexp substring repeat right char 5389 0 5000000
000 null uwep uwep',
 '4291 5023 ctxsys drithsx sn 5023 chr 113 chr 113 chr 112 chr 106 chr 113 select cas
e 5023 5023 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113 xyhi xyhi',
 '1 boolean mode 3707 select count sysibm systables t1 sysibm systables t2 sysibm sys
tables t3',
 '1 1022 select count user t1 user t2 user t3 user t4 user t5',
 'call regexp substring repeat left crypt key char 65 char 69 char 83 null 0 50000000
0 null pawh pawh',
 '1589 1589 1',
 'select count generate series 1 5000000 7240 7240',
 '1 3715 char 113 char 113 char 112 char 106 char 113 select case 3715 3715 char 49 e
lse char 48 end char 113 char 122 char 118 char 122 char 113 9548 9548',
 '1 6240 qqpjq select case 6240 6240 1 else 0 end rdb database qzvzq 6406 6406',
 '5021 select yadq 4285 4285 order 1',
 '1 4386 utl inaddr get host address chr 113 chr 113 chr 112 chr 106 chr 113 select c
ase 4386 4386 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113',
 '1 6793 select 6793 pg sleep 5 baia baia',
 '1 3824 benchmark 5000000 md5 0x76555642 jozh jozh',
 '9508 union select 9950 9950 9950 9950 9950 9950 9950',
 '1 8514 benchmark 5000000 md5 0x544d5a4c',
 '1064 make set 6439 2937 2937 qojd qojd',
 '1 rbpx 1264 1264',
 '1 exp select select concat 0x7171706a71 select elt 6270 6270 1 0x717a767a71 0x78
x',
 '8284 6171 select 8284 else drop function mbih',
 '1 sleep 5',
 '7868 9323 9323',
 '1 ekjw 5477 5477 union select null null null null null',
 '1 2462 2462 2716 select count sysusers sys1 sysusers sys2 sysusers sys3 sysusers sy
s4 sysusers sys5 sysusers sys6 sysusers sys7',
 'select like abcdefg upper hex randomblob 500000000 2 mfib mfib',
 '1 4240 select 4240 pg sleep 5',
 '1 4595 4595',
 '1 oknw 8777 8777',
 '1 6501 6501',
 '1 char 111 char 77 char 121 char 88 regexp substring repeat left crypt key char 65
char 69 char 83 null 0 500000000 null 8929 8929',
 '7999 8422 1336',
 '5742 1314 1314 5903 qqpjq select case 5903 5903 1 else 0 end rdb database qzvzq',
 '1 potk 5040 5040 elt 5873 5873 sleep 5',
 '4605 union select 8542 8542 8542 8542 8542 8542 8542 8542',
 'select count sysibm systables t1 sysibm systables t2 sysibm systables t3 njnr njn
r',
 'select case 6558 4327 1 else null end',
 '7535 2724 char 113 char 113 char 112 char 106 char 113 select case 2724 2724 char 4
9 else char 48 end char 113 char 122 char 118 char 122 char 113',
 'end rqay like rqay',
 '2312 union select 5282 5282 5282 5282 5282 5282 5282 5282',
 '1 8594 select 8594 pg sleep 5',
 'select case 8993 8846 8993 else 8993 select 8993 mysql db end',
 'select count sysibm systables t1 sysibm systables t2 sysibm systables t3',
 'select pg sleep 5',
 'call regexp substring repeat right char 3702 0 500000000 null 4142 4142',
 '8153 qhlb 4948 4948 union select 4948 4948 4948 4948 4948 4948',
 '7319 4493 utl inaddr get host address chr 113 chr 113 chr 112 chr 106 chr 113 selec
t case 4493 4493 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113',
 '1 bedq 8781 8781 8466 benchmark 5000000 md5 0x694a4745',
 '1 2006 2006',
 '5139 union select 3373 3373 3373 3373 3373 3373 3373 3373',
```

```
 '1 4411 select count sysusers sys1 sysusers sys2 sysusers sys3 sysusers sys4 sysuser
s sys5 sysusers sys6 sysusers sys7 zcyc',
 '1 rvch 1863 1863 8384 like abcdefg upper hex randomblob 500000000 2',
 '1 8514 select count domain domain t1 domain column t2 domain table t3 loao',
 '9446 wmrq 3705 3705 union select 3705',
 'iif 7889 5114 1 1 0',
 'select dbms pipe receive message chr 66 chr 67 chr 79 chr 101 5 dual ztmd ztmd',
 '1 union select null null null null null null null null',
 '1 char 109 char 79 char 70 char 90 regexp substring repeat right char 5012 0 500000
0000 null',
 '1 elt 4249 4249 7259',
 '1 4867 4867 rlike select case 7689 7689 1 else 0x28 end',
 '8858 5680 select 8858 else drop function pyuo',
 '1 procedure analyse extractvalue 9255 concat 0x5c benchmark 5000000 md5 0x52515a50
1 1748 1748',
 '1 klie 2840 2840 8514 benchmark 5000000 md5 0x544d5a4c',
 '3859 3440 cast chr 113 chr 113 chr 112 chr 106 chr 113 select case 3440 3440 1 else
0 end text chr 113 chr 122 chr 118 chr 122 chr 113 numeric 5846 5846',
 'end qkkn like qkkn',
 '1 pzoo 8036 8036 6793 select 6793 pg sleep 5',
 '2120 8734 8844',
 '1 lomw 9257 9257 union select null null null null null null',
 '1 6784 6784 elt 3114 3114 sleep 5',
 '1 select syrz 7699 7699 union select null null null',
 '1 4595 4595',
 '4984 union select 6980 6980 6980 6980 6980 6980 6980 6980',
 '5224 1962 1962 union select 1962 1962 1962 1962 1962 1962 1962 1962 1962',
 '1 8514 select count domain domain t1 domain column t2 domain table t3',
 '1 select gboi 4191 4191 8514 select count domain domain t1 domain column t2 domain
table t3',
 '1 union select null null null null null',
 '1 2633 dbms pipe receive message chr 112 chr 65 chr 65 chr 103 5 xmnd xmnd',
 '1 row 6237 7469 select count concat 0x7171706a71 select elt 6237 6237 1 0x717a767a7
1 floor rand 0 2 x select 5192 union select 3785 union select 3931 union select 7158
group x ejul ejul',
 '3721 union select 9050 9050',
 '1 paai 4089 4089 3707 select count sysibm systables t1 sysibm systables t2 sysibm s
ystables t3',
 '1 elt 3114 3114 sleep 5',
 '1 order 1',
 '9281 8363 8363 make set 8220 5127 5127',
 '1 6671 6671 char 119 char 100 char 99 char 121 regexp substring repeat right char 1
441 0 5000000000 null',
 'end vwbx vwbx',
 '1 boolean mode 8189 select count sysibm systables t1 sysibm systables t2 sysibm sys
tables t3',
 '1 rlike sleep 5 iwct iwct',
 '9087 order 1',
 '7562 8571 8571',
 '3707 8571 8571',
 '1 select twyt 3376 3376 7756 dbms utility sqlid sqlhash chr 113 chr 113 chr 112 chr
106 chr 113 select case 7756 7756 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 c
hr 113',
 'call regexp substring repeat right char 3702 0 500000000 null eevk eevk',
 '1 3754 select upper xmltype chr 60 chr 58 chr 113 chr 113 chr 112 chr 106 chr 113 s
elect case 3754 3754 1 else 0 end dual chr 113 chr 122 chr 118 chr 122 chr 113 chr 62
dual',
 ...]
```

In [80]:
```
data['Query'] = preprocessed_query
```

In [81]:
```
data.head()
```

Out[81]:

	Query	Label	no_sngle_quts	no_dble_quts	no_punctn	no_sgle_cmnt	no_mlt_cmnt	no_log_oprt
0	pg sleep time	1	0	1	10	1	0	
1	create user name identified pass123 temporary ...	1	0	0	1	0	0	
2	1 utl inaddr get host address select distinct ...	1	3	0	25	0	0	
3	select user id 1 1 1 union select 1 version 1	1	3	0	13	1	0	
4	select user id 1 1 union select 1 version 1	1	0	1	10	1	0	

5 rows × 23 columns