```
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
!pip install PyPDF2
import PyPDF2
if os.path.exists("kddcup.names.pdf"):
    with open("kddcup.names.pdf", 'rb') as f:
        pdf_reader = PyPDF2.PdfReader(f)
        for page_num in range(len(pdf_reader.pages)):
            page = pdf_reader.pages[page_num]
            print(page.extract_text())
else:
    print("File not found")



cols ="""duration,
protocol_type,
service,
flag,
src_bytes,
dst_bytes,
land,
wrong_fragment,
```

urgent,

hot,

num_failed_logins,

logged_in,

num_compromised,

root_shell,

su_attempted,

num_root,

num_file_creations,

num_shells,

num_access_files,

num_outbound_cmds,

is_host_login,

is_guest_login,

count,

srv_count,

serror_rate,

srv_serror_rate,

rerror_rate,

srv_rerror_rate,

same_srv_rate,

diff_srv_rate,

srv_diff_host_rate,

dst_host_count,

```
dst_host_srv_count,

dst_host_same_srv_rate,

dst_host_diff_srv_rate,

dst_host_same_src_port_rate,

dst_host_srv_diff_host_rate,

dst_host_serror_rate,

dst_host_srv_serror_rate,

dst_host_rerror_rate,

dst_host_srv_rerror_rate"""


columns =[]

for c in cols.split(',\n'):

   if(c.strip()):

      columns.append(c.strip())

columns.append('target')

print(len(columns))

with open("kdd.ics.uci.edudatabaseskddcup99training_attack_types.pdf", 'rb')
as f:

       pdf_reader = PyPDF2.PdfReader(f)

       for page_num in range(len(pdf_reader.pages)):

          page = pdf_reader.pages[page_num]

          print(page.extract_text())

attacks_types = {

   'normal': 'normal',

'back': 'dos',
```

'buffer_overflow': 'u2r',

'ftp_write': 'r2l',

'guess_passwd': 'r2l',

'imap': 'r2l',

'ipsweep': 'probe',

'land': 'dos',

'loadmodule': 'u2r',

'multihop': 'r2l',

'neptune': 'dos',

'nmap': 'probe',

'perl': 'u2r',

'phf': 'r2l',

'pod': 'dos',

'portsweep': 'probe',

'rootkit': 'u2r',

'satan': 'probe',

'smurf': 'dos',

'spy': 'r2l',

'teardrop': 'dos',

'warezclient': 'r2l',

'warezmaster': 'r2l',

}

path = "kddcup.data_10_percent.gz"

df = pd.read_csv(path, names = columns)

df['Attack Type'] = df.target.apply(lambda r:attacks_types[r[:-1]])

df.head(15)

df.shape

df.isnull().sum()

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

sns.countplot(x='Attack Type', data=df)

plt.xticks(rotation=90)

plt.title("Attack Type Distribution")

plt.show()

OUTPUT

42

back dos

buffer_overflow u2r

ftp_write r2l

guess_passwd r2l

imap r2l

ipsweep probe

land dos

loadmodule u2r

multihop r2l

neptune dos

nmap probe

perl u2r

phf r2l

pod dos

portsweep probe

rootkit u2r

satan probe

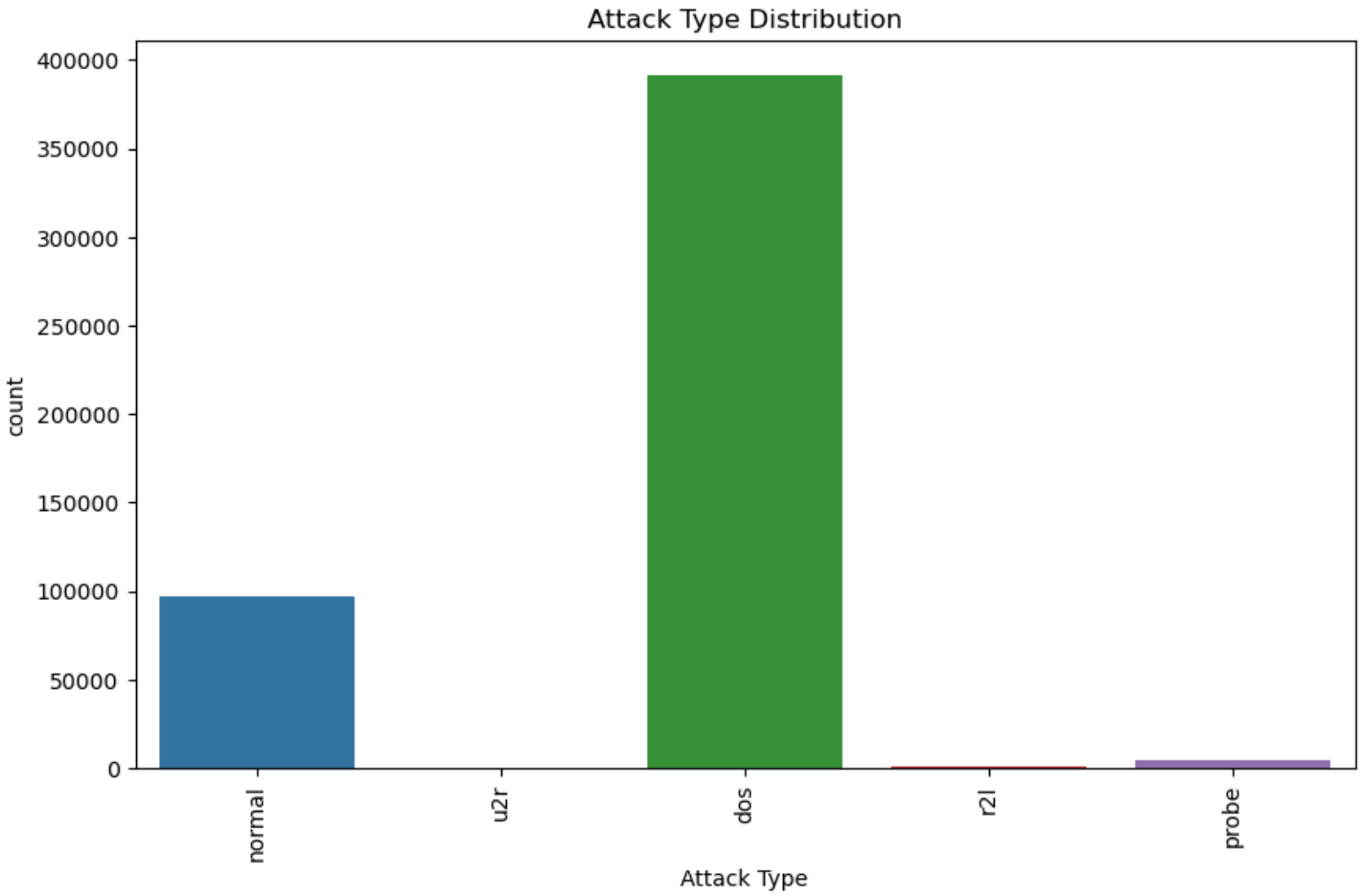
smurf dos

spy r2l

teardrop dos

warezclient r2l

warezmaster r2l



In [ ]: