

```
# I wasn't registered during this lab and so didn't had much idea as to what has to be done i
```

```
import pandas as pd
import numpy as np
```

```
data = pd.read_csv('KDD_Train.csv')
data.head()
```

```
↳
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	u
0	0	tcp	ftp_data	SF	491	0	0	0	
1	0	udp	other	SF	146	0	0	0	
2	0	tcp	private	S0	0	0	0	0	
3	0	tcp	http	SF	232	8153	0	0	
4	0	tcp	http	SF	199	420	0	0	

5 rows × 42 columns

+ Code

+ Text

```
data.isnull()
```

```

duration protocol_type service flag src_bytes dst_bytes land wrong_fragment
data.isnull().sum()

```

```

duration          0
protocol_type     0
service           0
flag              0
src_bytes         0
dst_bytes         0
land              0
wrong_fragment    0
urgent            0
hot               0
num_failed_logins 0
logged_in         0
num_compromised   0
root_shell        0
su_attempted      0
num_root          0
num_file_creations 0
num_shells        0
num_access_files  0
num_outbound_cmds 0
is_host_login     0
is_guest_login    0
count             0
srv_count         0
serror_rate       0
srv_serror_rate   0
rerror_rate       0
srv_rerror_rate   0
same_srv_rate     0
diff_srv_rate     0
srv_diff_host_rate 0
dst_host_count    0
dst_host_srv_count 0
dst_host_same_srv_rate 0
dst_host_diff_srv_rate 0
dst_host_same_src_port_rate 0
dst_host_srv_diff_host_rate 0
dst_host_serror_rate 0
dst_host_srv_serror_rate 0
dst_host_rerror_rate 0
dst_host_srv_rerror_rate 0
class             0
dtype: int64

```

```
data.duplicated()
```

```

0      False
1      False

```

```

2      False
3      False
4      False
...
125968 False
125969 False
125970 False
125971 False
125972 False
Length: 125973, dtype: bool

```

```
data.drop_duplicates()
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment
0	0	tcp	ftp_data	SF	491	0	0	
1	0	udp	other	SF	146	0	0	
2	0	tcp	private	S0	0	0	0	
3	0	tcp	http	SF	232	8153	0	
4	0	tcp	http	SF	199	420	0	
...	
125968	0	tcp	private	S0	0	0	0	
125969	8	udp	private	SF	105	145	0	
125970	0	tcp	smtp	SF	2231	384	0	
125971	0	tcp	klogin	S0	0	0	0	
125972	0	tcp	ftp_data	SF	151	0	0	

125964 rows × 42 columns

```
data['src_bytes'].describe()
```

```

count    1.259730e+05
mean     4.556674e+04
std      5.870331e+06
min      0.000000e+00
25%      0.000000e+00
50%      4.400000e+01
75%      2.760000e+02
max      1.379964e+09
Name: src_bytes, dtype: float64

```

[Colab paid products](#) - [Cancel contracts here](#)

