1.

In [1]:
```
import numpy as np
from numpy import random
```

In [2]:
```
bikeDataSet = np.genfromtxt('hour.csv', delimiter=',')

# Resources used:
# https://www.w3resource.com/numpy/input-and-output/genfromtxt.php
```

In [3]:
```
X = bikeDataSet[1:,2:-1]
y = bikeDataSet[1:,-1]
```

In [4]:
```
from sklearn import linear_model

estimator = linear_model.LinearRegression()
```

In [5]:
```
from sklearn.model_selection import cross_val_score

score = cross_val_score(estimator, X, y).mean()
print(f"Score = {score:.3f}.")
```

Score = 1.000.

In [6]:
```
X = np.random.rand(X.shape[0], 4)
```

In [7]:
```
score = cross_val_score(estimator, X, y).mean()
print(f"Score = {score:.3f}.")
```

Score = -0.280.

In [8]:
```
estimator = linear_model.Lasso(alpha = 0.1)

# Resources used:
# https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
```

In [9]:
```
score = cross_val_score(estimator, X, y).mean()
print(f"Score = {score:.3f}.")
```

Score = -0.280.

2.

In [10]:
```
import pandas as pd

df = pd.read_csv('kddcup.data', header = None)
```

In [11]:
```
df.columns = [
    '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',
    'outcome'
]
```

In [12]:
```
df.head()
```

Out[12]:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	...	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	outcome
0	0	tcp	http	SF	215	45076	0	0	0	0	...	0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	normal.
1	0	tcp	http	SF	162	4528	0	0	0	0	...	1	1.0	0.0	1.00	0.0	0.0	0.0	0.0	0.0	normal.
2	0	tcp	http	SF	236	1228	0	0	0	0	...	2	1.0	0.0	0.50	0.0	0.0	0.0	0.0	0.0	normal.
3	0	tcp	http	SF	233	2032	0	0	0	0	...	3	1.0	0.0	0.33	0.0	0.0	0.0	0.0	0.0	normal.
4	0	tcp	http	SF	239	486	0	0	0	0	...	4	1.0	0.0	0.25	0.0	0.0	0.0	0.0	0.0	normal.

5 rows × 42 columns

In [13]:
```
df.shape
```

Out[13]:
```
(4898431, 42)
```

In [14]:
```
df.dtypes
```

Out[14]:
```
duration                        int64
protocol_type                  object
service                        object
flag                           object
src_bytes                       int64
dst_bytes                       int64
land                            int64
wrong_fragment                  int64
urgent                          int64
hot                             int64
num_failed_logins               int64
logged_in                       int64
num_compromised                 int64
root_shell                      int64
su_attempted                    int64
num_root                        int64
num_file_creations              int64
num_shells                      int64
num_access_files                int64
num_outbound_cmds               int64
is_host_login                   int64
is_guest_login                  int64
count                           int64
srv_count                       int64
serror_rate                   float64
srv_serror_rate               float64
rerror_rate                   float64
srv_rerror_rate               float64
same_srv_rate                 float64
diff_srv_rate                 float64
srv_diff_host_rate            float64
dst_host_count                  int64
dst_host_srv_count              int64
dst_host_same_srv_rate        float64
dst_host_diff_srv_rate        float64
dst_host_same_src_port_rate   float64
dst_host_srv_diff_host_rate   float64
dst_host_serror_rate          float64
dst_host_srv_serror_rate      float64
dst_host_rerror_rate          float64
dst_host_srv_rerror_rate      float64
outcome                        object
dtype: object
```

In [15]:
```
df["protocol_type"] = pd.Categorical(df["protocol_type"])
df["service"] = pd.Categorical(df["service"])
df["flag"] = pd.Categorical(df["flag"])
df["outcome"] = pd.Categorical(df["outcome"])

# Resources used:
# https://www.geeksforgeeks.org/python-pandas-categorical/#
```

In [16]:
```
df = pd.get_dummies(df, columns = ['protocol_type', 'service', 'flag', 'outcome'])
```

In [17]:
```
df.head()
```

Out[17]:

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	...	outcome_phf.	outcome_pod.	outcome_portsweep.	outcome_rootkit.	outcome_satan.	outcome_smurf.	outcome_spy.	outcome_teardrop.	outcome_warezclient.	outcome_warezmaster.
0	0	215	45076	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
1	0	162	4528	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
2	0	236	1228	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
3	0	233	2032	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
4	0	239	486	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 145 columns

In [18]:
```
df.shape
```

Out[18]:
```
(4898431, 145)
```

In [19]:
```
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
```

In [20]:
```
norm = df.columns[:38]
```

In [21]:
```
df[norm] = sc.fit_transform(df[norm])
```

In [22]:
```
df[norm].mean
```

Out[22]:
```
<bound method NDFrame._add_numeric_operations.<locals>.mean of          duration  src_bytes  dst_bytes      land  wrong_fragment    urgent  \
0       -0.006833 -0.001720  0.068188 -0.002391       -0.015139 -0.001103
1       -0.006833 -0.001777  0.005325 -0.002391       -0.015139 -0.001103
2       -0.006833 -0.001698  0.000208 -0.002391       -0.015139 -0.001103
3       -0.006833 -0.001701  0.001455 -0.002391       -0.015139 -0.001103
4       -0.006833 -0.001695 -0.000942 -0.002391       -0.015139 -0.001103
...           ...       ...       ...       ...             ...       ...
4898426 -0.006833 -0.001724  0.001052 -0.002391       -0.015139 -0.001103
4898427 -0.006833 -0.001716 -0.001338 -0.002391       -0.015139 -0.001103
4898428 -0.006833 -0.001717  0.003901 -0.002391       -0.015139 -0.001103
4898429 -0.006833 -0.001716  0.000218 -0.002391       -0.015139 -0.001103
4898430 -0.006833 -0.001716  0.000007 -0.002391       -0.015139 -0.001103

              hot  num_failed_logins  logged_in  num_compromised  ...  \
0       -0.026521          -0.004391   2.442792        -0.002097  ...
1       -0.026521          -0.004391   2.442792        -0.002097  ...
2       -0.026521          -0.004391   2.442792        -0.002097  ...
3       -0.026521          -0.004391   2.442792        -0.002097  ...
4       -0.026521          -0.004391   2.442792        -0.002097  ...
...           ...                ...        ...             ...  ...
4898426 -0.026521          -0.004391   2.442792        -0.002097  ...
4898427 -0.026521          -0.004391   2.442792        -0.002097  ...
4898428 -0.026521          -0.004391   2.442792        -0.002097  ...
4898429 -0.026521          -0.004391   2.442792        -0.002097  ...
4898430 -0.026521          -0.004391   2.442792        -0.002097  ...

         dst_host_count  dst_host_srv_count  dst_host_same_srv_rate  \
0             -3.639139           -1.786510               -1.833023
1             -3.623519           -1.777069                0.598987
2             -3.607899           -1.767627                0.598967
3             -3.592280           -1.758185                0.598967
4             -3.576660           -1.748744                0.598967
...                 ...                 ...                     ...
4898426       -1.552230            0.621132                0.598967
4898427       -1.576660            0.621132                0.598967
4898428       -1.561040            0.621132                0.598967
4898429       -1.545420            0.621132                0.598967
4898430       -1.529800            0.621132                0.598967

         dst_host_diff_srv_rate  dst_host_same_src_port_rate  \
0                    -0.282939                    -1.257937
1                    -0.282939                     0.821119
2                    -0.282939                    -0.218409
3                    -0.282939                    -0.571848
4                    -0.282939                    -0.738173
...                        ...                          ...
4898426              -0.282939                    -0.571848
4898427              -0.282939                    -0.738173
4898428              -0.282939                    -0.842126
4898429              -0.282939                    -0.904407
4898430              -0.282939                    -0.966869

         dst_host_srv_diff_host_rate  dst_host_serror_rate  \
0                          -0.156668             -0.466405
1                          -0.156668             -0.466405
2                          -0.156668             -0.466405
3                          -0.156668             -0.466405
4                          -0.156668             -0.466405
...                              ...                   ...
4898426                     1.055166             -0.466405
4898427                     1.055166             -0.466405
4898428                     1.055166             -0.466405
4898429                     1.055166             -0.466405
4898430                     1.055166             -0.466405

         dst_host_srv_serror_rate  dst_host_rerror_rate  \
0                       -0.465454             -0.250832
1                       -0.465454             -0.250832
2                       -0.465454             -0.250832
3                       -0.465454             -0.250832
4                       -0.465454             -0.250832
...                           ...                   ...
4898426                 -0.439288             -0.250832
4898427                 -0.439288             -0.250832
4898428                 -0.439288             -0.250832
4898429                 -0.439288             -0.250832
4898430                 -0.439288             -0.250832

         dst_host_srv_rerror_rate
0                       -0.249632
1                       -0.249632
2                       -0.249632
3                       -0.249632
4                       -0.249632
...                           ...
4898426                 -0.249632
4898427                 -0.249632
4898428                 -0.249632
4898429                 -0.249632
4898430                 -0.249632

[4898431 rows x 38 columns]>
```

In [23]:
```
df[norm].std()
```

Out[23]:
```
duration                        1.0
src_bytes                       1.0
dst_bytes                       1.0
land                            1.0
wrong_fragment                  1.0
urgent                          1.0
hot                             1.0
num_failed_logins               1.0
logged_in                       1.0
num_compromised                 1.0
root_shell                      1.0
su_attempted                    1.0
num_root                        1.0
num_file_creations              1.0
num_shells                      1.0
num_access_files                1.0
num_outbound_cmds               0.0
is_host_login                   1.0
is_guest_login                  1.0
count                           1.0
srv_count                       1.0
serror_rate                     1.0
srv_serror_rate                 1.0
rerror_rate                     1.0
srv_rerror_rate                 1.0
same_srv_rate                   1.0
diff_srv_rate                   1.0
srv_diff_host_rate              1.0
dst_host_count                  1.0
dst_host_srv_count              1.0
dst_host_same_srv_rate          1.0
dst_host_diff_srv_rate          1.0
dst_host_same_src_port_rate     1.0
dst_host_srv_diff_host_rate     1.0
dst_host_serror_rate            1.0
dst_host_srv_serror_rate        1.0
dst_host_rerror_rate            1.0
dst_host_srv_rerror_rate        1.0
dtype: float64
```

In [24]:
```
from sklearn import datasets
from scipy.stats import describe
from sklearn.svm import SVC
```

3.

In [25]:
```
alpha = 100000
```

In [26]:
```
iris = datasets.load_iris()

X = iris.data
y = iris.target
```

In [27]:
```
X = X * alpha
```

In [28]:
```
estimator = SVC(kernel = 'linear')
score = cross_val_score(estimator, X, y).mean()
print(f"Score = {score:.3f}.")
```

Score = 0.973.