In [2]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Put this when it's called
from sklearn.model_selection import train_test_split
from sklearn.model_selection import learning_curve
from sklearn.model_selection import validation_curve
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
```

In [3]:

```
def draw_missing_data_table(df):
    total = df.isnull().sum().sort_values(ascending=False)
    percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
    missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
    return missing_data
```

In [4]:

```
def plot_learning_curve(estimator, title, X, y, ylim=None, cv=None,
                        n_jobs=1, train_sizes=np.linspace(.1, 1.0, 5)):
    plt.figure()
    plt.title(title)
    if ylim is not None:
        plt.ylim(*ylim)
    plt.xlabel("Training examples")
    plt.ylabel("Score")
    train_sizes, train_scores, test_scores = learning_curve(
        estimator, X, y, cv=cv, n_jobs=n_jobs, train_sizes=train_sizes)
    train_scores_mean = np.mean(train_scores, axis=1)
    train_scores_std = np.std(train_scores, axis=1)
    test_scores_mean = np.mean(test_scores, axis=1)
    test_scores_std = np.std(test_scores, axis=1)
    plt.grid()

    plt.fill_between(train_sizes, train_scores_mean - train_scores_std,
                     train_scores_mean + train_scores_std, alpha=0.1,
                     color="r")
    plt.fill_between(train_sizes, test_scores_mean - test_scores_std,
                     test_scores_mean + test_scores_std, alpha=0.1, color="g")
    plt.plot(train_sizes, train_scores_mean, 'o-', color="r",
             label="Training score")
    plt.plot(train_sizes, test_scores_mean, 'o-', color="g",
             label="Validation score")

    plt.legend(loc="best")
    return plt
```

In [5]:

```
def plot_validation_curve(estimator, title, X, y, param_name, param_range, ylim=None, cv=None,
                          n_jobs=1, train_sizes=np.linspace(.1, 1.0, 5)):
    train_scores, test_scores = validation_curve(estimator, X, y, param_name, param_range, cv)
    train_mean = np.mean(train_scores, axis=1)
    train_std = np.std(train_scores, axis=1)
    test_mean = np.mean(test_scores, axis=1)
    test_std = np.std(test_scores, axis=1)
    plt.plot(param_range, train_mean, color='r', marker='o', markersize=5, label='Training score')
    plt.fill_between(param_range, train_mean + train_std, train_mean - train_std, alpha=0.15, color='r')
    plt.plot(param_range, test_mean, color='g', linestyle='--', marker='s', markersize=5, label='Validation score')
    plt.fill_between(param_range, test_mean + test_std, test_mean - test_std, alpha=0.15, color='g')
    plt.grid()
    plt.xscale('log')
    plt.legend(loc='best')
    plt.xlabel('Parameter')
    plt.ylabel('Score')
    plt.ylim(ylim)
```

In [6]:

```
df = pd.read_csv(r'C:\Users\Anustup\Desktop\dataset_malwares.csv')
df_raw = df.copy()
```

In [7]:

```
df.head()
```

Out[7]:

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr
0	VirusShare_a878ba26000edaac5c98eff4432723b3	23117	144	3	0	4
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	23117	144	3	0	4
2	VirusShare_ef84cdeba22be72a69b198213dada81a	23117	144	3	0	4
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	23117	144	3	0	4
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	23117	144	3	0	4

5 rows × 79 columns

In [8]:

```
df.describe()
```

Out[8]:

	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc	e_
count	19611.0	19611.000000	19611.000000	19611.000000	19611.000000	19611.000000	1961
mean	23117.0	178.615726	71.660752	49.146958	37.370710	37.032635	6417
std	0.0	987.200729	1445.192977	1212.201919	864.515405	915.833139	911
min	23117.0	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	23117.0	144.000000	3.000000	0.000000	4.000000	0.000000	6553
50%	23117.0	144.000000	3.000000	0.000000	4.000000	0.000000	6553
75%	23117.0	144.000000	3.000000	0.000000	4.000000	0.000000	6553
max	23117.0	59448.000000	63200.000000	64613.000000	43690.000000	43690.000000	6553

8 rows × 78 columns

In [9]:

```
draw_missing_data_table(df)
```

Out[9]:

	Total	Percent
ImageDirectoryEntrySecurity	0	0.0
SizeOfCode	0	0.0
PointerToSymbolTable	0	0.0
NumberOfSymbols	0	0.0
SizeOfOptionalHeader	0	0.0
Characteristics	0	0.0
Magic	0	0.0
MajorLinkerVersion	0	0.0
MinorLinkerVersion	0	0.0
SizeOfInitializedData	0	0.0
NumberOfSections	0	0.0
SizeOfUninitializedData	0	0.0
AddressOfEntryPoint	0	0.0
BaseOfCode	0	0.0
ImageBase	0	0.0
SectionAlignment	0	0.0
FileAlignment	0	0.0
MajorOperatingSystemVersion	0	0.0
TimeDateStamp	0	0.0
Machine	0	0.0
MajorImageVersion	0	0.0
e_ss	0	0.0
e_magic	0	0.0
e_cblp	0	0.0
e_cp	0	0.0
e_crlc	0	0.0
e_cparhdr	0	0.0
e_minalloc	0	0.0
e_maxalloc	0	0.0
e_sp	0	0.0
...	...	...
SectionMaxChar	0	0.0
SectionMinRawsize	0	0.0
SectionMainChar	0	0.0
DirectoryEntryImport	0	0.0
DirectoryEntryImportSize	0	0.0
DirectoryEntryExport	0	0.0

	Total	Percent
ImageDirectoryEntryExport	0	0.0
ImageDirectoryEntryImport	0	0.0
ImageDirectoryEntryResource	0	0.0
SectionMaxRawsize	0	0.0
SectionMaxEntropy	0	0.0
MajorSubsystemVersion	0	0.0
SizeOfStackCommit	0	0.0
MinorSubsystemVersion	0	0.0
SizeOfHeaders	0	0.0
CheckSum	0	0.0
SizeOfImage	0	0.0
Subsystem	0	0.0
DllCharacteristics	0	0.0
SizeOfStackReserve	0	0.0
SizeOfHeapReserve	0	0.0
SectionMinEntropy	0	0.0
SizeOfHeapCommit	0	0.0
LoaderFlags	0	0.0
NumberOfRvaAndSizes	0	0.0
Malware	0	0.0
SuspiciousImportFunctions	0	0.0
SuspiciousNameSection	0	0.0
SectionsLength	0	0.0
Name	0	0.0

79 rows × 2 columns

In [10]:

```
df.drop('e_magic', axis=1, inplace=True)
df.head()
```

Out[10]:

	Name	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc
0	VirusShare_a878ba26000edaac5c98eff4432723b3	144	3	0	4	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	144	3	0	4	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	144	3	0	4	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	144	3	0	4	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	144	3	0	4	0

5 rows × 78 columns

In [11]:

```
value = 1000
df['e_cp'].fillna(1000, inplace=True)
df['e_cp'].max()
```

Out[11]:

```
63200
```

In [12]:

```
df.drop(df[pd.isnull(df['e_cblp'])].index, inplace=True)
df[pd.isnull(df['e_cblp'])]
```

Out[12]:

	Name	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc	e_maxalloc	e_ss	e_sp	e_csum	...

0 rows × 78 columns

In [13]:

```
df.drop('e_cblp', axis=1, inplace=True)
df.head()
```

Out[13]:

	Name	e_cp	e_crlc	e_cparhdr	e_minalloc	e_max
0	VirusShare_a878ba26000edaac5c98eff4432723b3	3	0	4	0	6
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	3	0	4	0	6
2	VirusShare_ef84cdeba22be72a69b198213dada81a	3	0	4	0	6
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	3	0	4	0	6
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	3	0	4	0	6

5 rows × 77 columns

In [14]:

```
df['e_cp'] = pd.Categorical(df['e_cp'])
df['e_crlc'] = pd.Categorical(df['e_crlc'])
```

In [15]:

```
df.drop('e_cp',axis=1,inplace=True)
df.drop('e_crlc',axis=1,inplace=True)
df.head()
```

Out[15]:

	Name	e_cparhdr	e_minalloc	e_maxalloc	e_ss	e
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	0	65535	0	
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	0	65535	0	
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	0	65535	0	
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	0	65535	0	
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	0	65535	0	

5 rows × 75 columns

In [16]:

```
# Drop Name and Ticket
df.drop('e_cparhdr', axis=1, inplace=True)
df.drop('e_minalloc', axis=1, inplace=True)
df.head()
```

Out[16]:

	Name	e_maxalloc	e_ss	e_sp	e_csum	e_ip	e_c
0	VirusShare_a878ba26000edaac5c98eff4432723b3	65535	0	184	0	0	
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	65535	0	184	0	0	
2	VirusShare_ef84cdeba22be72a69b198213dada81a	65535	0	184	0	0	
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	65535	0	184	0	0	
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	65535	0	184	0	0	

5 rows × 73 columns

In [17]:

```
df = pd.get_dummies(df, drop_first=True)
df.head()
```

Out[17]:

	e_maxalloc	e_ss	e_sp	e_csum	e_ip	e_cs	e_lfarlc	e_ovno	e_oemid	e_oeminfo	...
0	65535	0	184	0	0	0	64	0	0	0	...
1	65535	0	184	0	0	0	64	0	0	0	...
2	65535	0	184	0	0	0	64	0	0	0	...
3	65535	0	184	0	0	0	64	0	0	0	...
4	65535	0	184	0	0	0	64	0	0	0	...

5 rows × 19682 columns

In [18]:

```
# Create data set to train data imputation methods
X = df[df.loc[:, df.columns != 'e_maxalloc'].columns]
y = df['e_maxalloc']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.2, random_state=1)
```

In [19]:

```
print('Inputs: \n', X_train.head())
print('Outputs: \n', y_train.head())
```

```
Inputs:
        e_ss  e_sp  e_csum  e_ip  e_cs  e_lfarlc  e_ovno  e_oemid  e_oemin
fo  \
17893     0   184       0     0     0        64       0        0
0
5886      0   184       0     0     0        64      26        0
0
17817     0   184       0     0     0        64       0        0
0
5511      0   184       0     0     0        64      26        0
0
2706      0   184       0     0     0        64       0        0
0

       e_lfanew        ...         Name_xrXpsPtFilter.DLL  Name_xul.dll  \
17893       248        ...                              0             0
5886        256        ...                              0             0
17817       232        ...                              0             0
5511        256        ...                              0             0
2706        128        ...                              0             0

       Name_xwizard.exe  Name_xwizards.dll  Name_xwreg.dll  Name_xwtpdui.d
ll  \
17893                 0                  0               0
0
5886                  0                  0               0
0
17817                 0                  0               0
0
5511                  0                  0               0
0
2706                  0                  0               0
0

       Name_xwtpw32.dll  Name_yara.dll  Name_zip.dll  Name_zipfldr.dll
17893                 0              0             0                 0
5886                  0              0             0                 0
17817                 0              0             0                 1
5511                  0              0             0                 0
2706                  0              0             0                 0

[5 rows x 19681 columns]
Outputs:
 17893    65535
5886     65535
17817    65535
5511     65535
2706     65535
Name: e_maxalloc, dtype: int64
```

In [20]:

```
# Fit logistic regression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
```

Out[20]:

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=Tru
e,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

In [21]:

```
scores = cross_val_score(logreg, X_train, y_train, cv=10)
print('CV accuracy: %.3f +/- %.3f' % (np.mean(scores), np.std(scores)))
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\model_selection\_spli
t.py:652: Warning: The least populated class in y has only 1 members, whic
h is too few. The minimum number of members in any class cannot be less th
an n_splits=10.
  % (min_groups, self.n_splits)), Warning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
```

```
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)


CV accuracy: 0.983 +/- 0.003
```

In [22]:

```
# Plot learning curves
title = "Learning Curves (Logistic Regression)"
cv = 10
plot_learning_curve(logreg, title, X_train, y_train, ylim=(0.7, 1.01), cv=cv, n_jobs=1
);
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\model_selection\_spli
t.py:652: Warning: The least populated class in y has only 1 members, whic
h is too few. The minimum number of members in any class cannot be less th
an n_splits=10.
  % (min_groups, self.n_splits)), Warning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
```

```
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
```

```
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
```

```
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
```

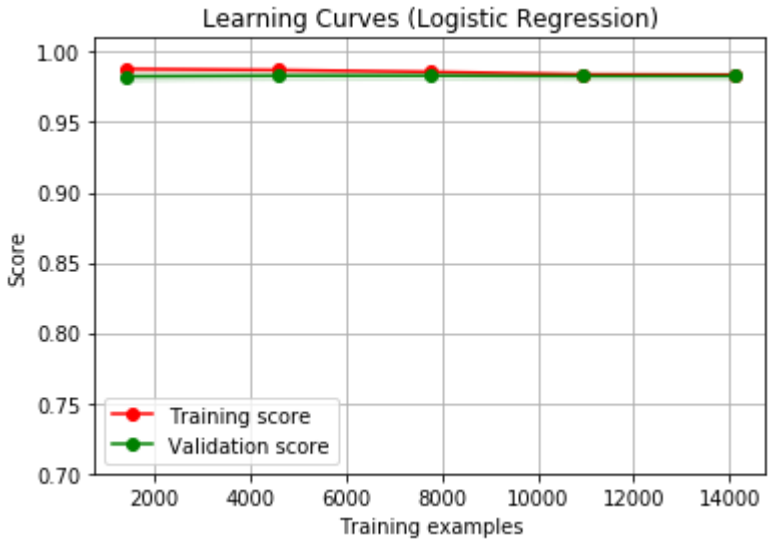
```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
```

```
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
```

```
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
```

```
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
```


Learning Curves (Logistic Regression)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Score
Training score
Validation score
2000
4000
6000
8000
10000
12000
14000
Training examples

In [23]:

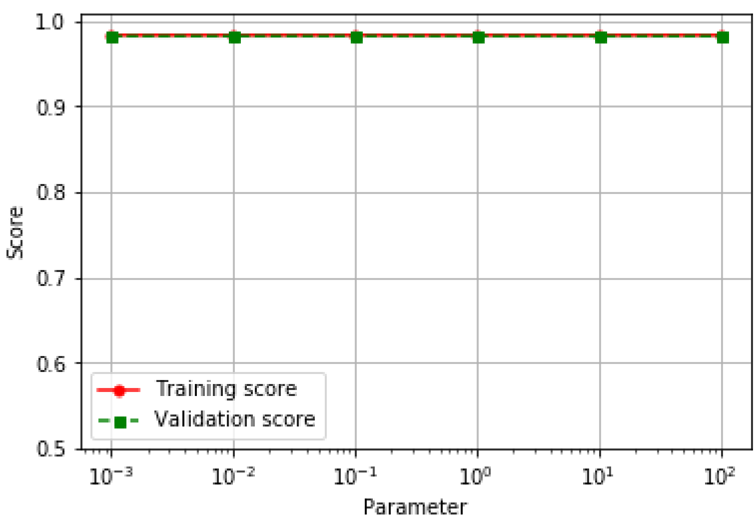
```
# Plot validation curve
title = 'Validation Curve (Logistic Regression)'
param_name = 'C'
param_range = [0.001, 0.01, 0.1, 1.0, 10.0, 100.0]
cv = 10
plot_validation_curve(estimator=logreg, title=title, X=X_train, y=y_train, param_name=p
aram_name,
                      ylim=(0.5, 1.01), param_range=param_range);
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\model_selection\_spli
t.py:2053: FutureWarning: You should specify a value for 'cv' instead of r
elying on the default value. The default value will change from 3 to 5 in
version 0.22.
  warnings.warn(CV_WARNING, FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\model_selection\_spli
t.py:652: Warning: The least populated class in y has only 1 members, whic
h is too few. The minimum number of members in any class cannot be less th
an n_splits=3.
  % (min_groups, self.n_splits)), Warning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
```

```
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
```

```
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
```



In [24]:

```
# Restart data set
df = df_raw.copy()
df.head()
```

Out[24]:

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr
0	VirusShare_a878ba26000edaac5c98eff4432723b3	23117	144	3	0	4
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	23117	144	3	0	4
2	VirusShare_ef84cdeba22be72a69b198213dada81a	23117	144	3	0	4
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	23117	144	3	0	4
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	23117	144	3	0	4

5 rows × 79 columns

In [25]:

```
df.drop('e_magic',axis=1,inplace=True)
df.drop('e_cblp',axis=1,inplace=True)
df.head()
```

Out[25]:

	Name	e_cp	e_crlc	e_cparhdr	e_minalloc	e_max
0	VirusShare_a878ba26000edaac5c98eff4432723b3	3	0	4	0	6
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	3	0	4	0	6
2	VirusShare_ef84cdeba22be72a69b198213dada81a	3	0	4	0	6
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	3	0	4	0	6
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	3	0	4	0	6

5 rows × 77 columns

In [26]:

```
# Drop irrelevant features
df.drop(['e_cp','e_crlc','e_minalloc'], axis=1, inplace=True)
df.head()
```

Out[26]:

	Name	e_cparhdr	e_maxalloc	e_ss	e_sp	e_csum
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	65535	0	184	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	65535	0	184	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	65535	0	184	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	65535	0	184	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	65535	0	184	0

5 rows × 74 columns

In [27]:

```
df_raw['Name'].unique()[:10]
```

Out[27]:

```
array(['VirusShare_a878ba26000edaac5c98eff4432723b3',
       'VirusShare_ef9130570fddc174b312b2047f5f4cf0',
       'VirusShare_ef84cdeba22be72a69b198213dada81a',
       'VirusShare_6bf3608e60ebc16cbcff6ed5467d469e',
       'VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb',
       'VirusShare_eff7676f69be2b519f3424def92d3590',
       'VirusShare_e76cac211258723745f66bd9f9e29590',
       'VirusShare_cef6cdf0e85303a461f67f19ffcc2ddf',
       'VirusShare_59af5dfb0c79537eedd3326abde3c857',
       'VirusShare_fda0add9d9a8c18c67a758ec2898d976'], dtype=object)
```

In [28]:

```
for i in df:
    df['Title']=df_raw['Name'].str.extract('([A-Za-z]+)\.', expand=False)  # Use REGEX
 to define a search pattern
df.head()
```

Out[28]:

	Name	e_cparhdr	e_maxalloc	e_ss	e_sp	e_csum
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	65535	0	184	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	65535	0	184	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	65535	0	184	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	65535	0	184	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	65535	0	184	0

5 rows × 75 columns

In [29]:

```
df_raw['Name'].unique()[:10]
```

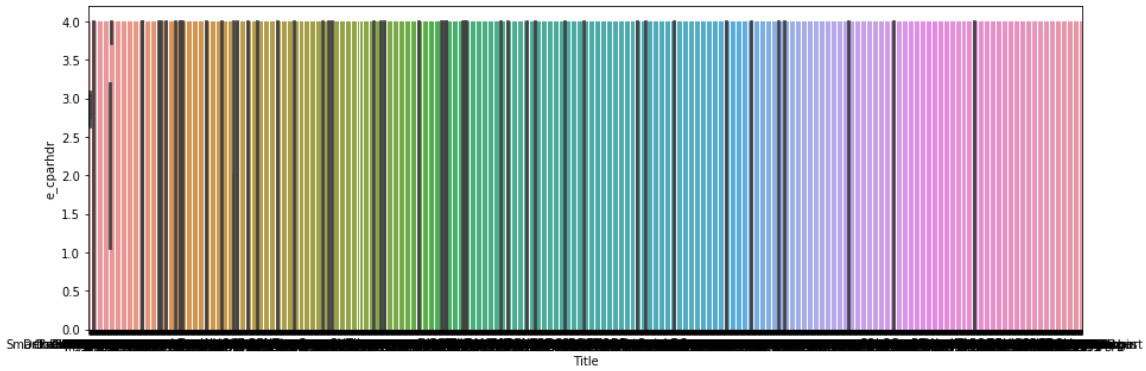
Out[29]:

```
array(['VirusShare_a878ba26000edaac5c98eff4432723b3',
       'VirusShare_ef9130570fddc174b312b2047f5f4cf0',
       'VirusShare_ef84cdeba22be72a69b198213dada81a',
       'VirusShare_6bf3608e60ebc16cbcff6ed5467d469e',
       'VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb',
       'VirusShare_eff7676f69be2b519f3424def92d3590',
       'VirusShare_e76cac211258723745f66bd9f9e29590',
       'VirusShare_cef6cdf0e85303a461f67f19ffcc2ddf',
       'VirusShare_59af5dfb0c79537eedd3326abde3c857',
       'VirusShare_fda0add9d9a8c18c67a758ec2898d976'], dtype=object)
```

In [30]:

```
# Plot bar plot (titles, age and sex)
plt.figure(figsize=(15,5))
sns.barplot(x=df['Title'], y=df_raw['e_cparhdr']);
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Fu
tureWarning: Using a non-tuple sequence for multidimensional indexing is d
eprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this
will be interpreted as an array index, `arr[np.array(seq)]`, which will re
sult either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



In [31]:

```
# Means per title
df_raw['Title'] = df['Title']  # To simplify data handling
means = df_raw.groupby('Title')['e_cparhdr'].mean()
means.head()
```

Out[31]:

```
Title
A           4.0
ACCTRES     4.0
ARP         4.0
AUDIOKSE    4.0
AagMmcRes   4.0
Name: e_cparhdr, dtype: float64
```

In [32]:

```
# Transform means into a dictionary for future mapping
map_means = means.to_dict()
map_means
```

Out[32]:


```
{'A': 4.0,
 'ACCTRES': 4.0,
 'ARP': 4.0,
 'AUDIOKSE': 4.0,
 'AagMmcRes': 4.0,
 'AcGenral': 4.0,
 'AcLayers': 4.0,
 'AcRes': 4.0,
 'AcWinRT': 4.0,
 'AcXtrnal': 4.0,
 'Accessibility': 2.0,
 'AccessibleHandler': 4.0,
 'AccessibleMarshal': 4.0,
 'ActionCenter': 4.0,
 'ActionCenterCPL': 4.0,
 'ActionQueue': 4.0,
 'ActiveDirectoryPowerShellResources': 4.0,
 'ActiveSockets': 4.0,
 'AddInProcess': 4.0,
 'AddInUtil': 4.0,
 'AdmTmpl': 4.0,
 'AdoNetDiag': 4.0,
 'AepRoam': 4.0,
 'Agent': 4.0,
 'Alduin': 4.0,
 'AltTab': 4.0,
 'AppHostNavigators': 4.0,
 'AppIdPolicyEngineApi': 4.0,
 'AppLaunch': 4.0,
 'AppReadiness': 4.0,
 'AppVStreamingUX': 4.0,
 'AppXDeploymentClient': 4.0,
 'AppXDeploymentExtensions': 4.0,
 'AppXDeploymentServer': 4.0,
 'Apphlpdm': 4.0,
 'AppxAllUserStore': 4.0,
 'AppxApplicabilityEngine': 4.0,
 'AppxPackaging': 4.0,
 'AppxProvider': 4.0,
 'AppxSip': 4.0,
 'AppxStreamingDataSourcePS': 4.0,
 'AppxSysprep': 4.0,
 'AppxUpgradeMigrationPlugin': 4.0,
 'ArpCacheWatch': 4.0,
 'AssocProvider': 4.0,
 'AtBroker': 4.0,
 'AudioEndpointBuilder': 4.0,
 'AudioEng': 4.0,
 'AudioSes': 4.0,
 'AuditNativeSnapIn': 4.0,
 'AuditPolicyGPInterop': 4.0,
 'AuditPolicyGPManagedStubs': 2.0,
 'AuditShD': 4.0,
 'AuthBroker': 4.0,
 'AuthExt': 4.0,
 'AuthFWGP': 4.0,
 'AuthFWSnapIn': 4.0,
 'AuthFWSnapin': 4.0,
 'AuthFWWizFwk': 4.0,
```

```
    'AuthHost': 4.0,
    'AuthHostProxy': 4.0,
    'AuthenticAMD': 4.0,
    'AutoWorkplace': 4.0,
    'AutoWorkplaceN': 4.0,
    'AzSqlExt': 4.0,
    'B': 4.0,
    'BEasyDiscBurner': 4.0,
    'BFE': 4.0,
    'BGPCore': 4.0,
    'BOOKMARK': 4.0,
    'BOOTVID': 4.0,
    'BPAInst': 4.0,
    'BWContextHandler': 4.0,
    'BackgroundTransferHost': 4.0,
    'BatchParser': 4.0,
    'Battle': 4.0,
    'Bennett': 4.0,
    'BgpRasMgmt': 4.0,
    'BlackBirdImageOptimizer': 4.0,
    'BlbEvents': 4.0,
    'BlizzardError': 4.0,
    'BootMenuUX': 4.0,
    'BthMtpContextHandler': 4.0,
    'ByteCodeGenerator': 4.0,
    'C': 4.0,
    'CHxReadingStringIME': 4.0,
    'CNBJOPAD': 4.0,
    'CNBJOPAE': 4.0,
    'CNBJOPAG': 4.0,
    'CNBJOPAI': 4.0,
    'CNBJOPAQ': 4.0,
    'CNBJOPAT': 4.0,
    'CNBJOPAU': 4.0,
    'CORPerfMonExt': 4.0,
    'CSDeployRes': 4.0,
    'CSystemEventsBrokerClient': 4.0,
    'CallButtons': 4.0,
    'CasPol': 4.0,
    'CbsApi': 4.0,
    'CbsCore': 4.0,
    'CbsMsg': 4.0,
    'CbsProvider': 4.0,
    'CertEnroll': 4.0,
    'CertEnrollCtrl': 4.0,
    'CertEnrollUI': 4.0,
    'CertPolEng': 4.0,
    'CheckNetIsolation': 4.0,
    'ChsIME': 4.0,
    'ChsPinyinDS': 4.0,
    'ChsRoaming': 4.0,
    'ChsWubiDS': 4.0,
    'ChtChangjieDS': 4.0,
    'ChtIME': 4.0,
    'ChtPhoneticDS': 4.0,
    'ChtQuickDS': 4.0,
    'ChxAdvancedDS': 4.0,
    'ChxEM': 4.0,
    'ChxProxyDS': 4.0,
    'ChxUserDictDS': 4.0,
    'CimProvider': 4.0,
```

```
    'ClientSdk': 4.0,
    'ClientSdkFirewallHelper': 4.0,
    'ClientSdkMDNSHost': 4.0,
    'Cmdline': 4.0,
    'CntrtextInstaller': 4.0,
    'CntrtextMig': 4.0,
    'ComSvcConfig': 2.6666666666666665,
    'CompMgmtLauncher': 4.0,
    'CompPkgSup': 4.0,
    'Compat': 4.0,
    'CompatProvider': 4.0,
    'ComplianceExtensions': 4.0,
    'ComputerDefaults': 4.0,
    'ConfigSvc': 4.0,
    'ConfigureExpandedStorage': 4.0,
    'ConfigureIEOptionalComponentsAI': 4.0,
    'Conn': 4.0,
    'ConnectedAccountState': 4.0,
    'Core': 4.0,
    'CoreMmRes': 4.0,
    'CoreNetHealth': 4.0,
    'CredentialUIBroker': 4.0,
    'CryptoWinRT': 4.0,
    'Cubby': 4.0,
    'Culture': 4.0,
    'CustomMarshalers': 2.6666666666666665,
    'CvtResUI': 4.0,
    'D': 4.0,
    'DAConn': 4.0,
    'DAFWSD': 4.0,
    'DAMM': 4.0,
    'DAMigPlugin': 4.0,
    'DBus': 4.0,
    'DDOIProxy': 4.0,
    'DDORes': 4.0,
    'DHCPQEC': 4.0,
    'DIFxAPI': 4.0,
    'DSETUP': 4.0,
    'DU': 4.0,
    'DWWIN': 4.0,
    'DWrite': 4.0,
    'DXP': 4.0,
    'DXSETUP': 4.0,
    'DaOtpAuth': 4.0,
    'DafPrintProvider': 4.0,
    'DataSvcUtil': 4.0,
    'DefaultDeviceManager': 4.0,
    'DefaultPrinterProvider': 4.0,
    'Defrag': 4.0,
    'Detect': 4.0,
    'DevDispItemProvider': 4.0,
    'DevPropMgr': 4.0,
    'DeviceCenter': 4.0,
    'DeviceDisplayStatusManager': 4.0,
    'DeviceDriverRetrievalClient': 4.0,
    'DeviceEject': 4.0,
    'DeviceElementSource': 4.0,
    'DeviceMetadataRetrievalClient': 4.0,
    'DeviceNameResolver': 4.0,
    'DevicePairing': 4.0,
    'DevicePairingFolder': 4.0,
```

```
    'DevicePairingProxy': 4.0,
    'DevicePairingWizard': 4.0,
    'DeviceProperties': 4.0,
    'DeviceSetupManager': 4.0,
    'DeviceSetupManagerAPI': 4.0,
    'DeviceSetupStatusProvider': 4.0,
    'DeviceUxRes': 4.0,
    'DfsDiag': 4.0,
    'DfsRes': 4.0,
    'DfsShlEx': 4.0,
    'DfsrHelper': 4.0,
    'DhcpServerPSProvider': 4.0,
    'DiagCpl': 4.0,
    'DiagPackage': 4.0,
    'DiagnosticsHub': 4.0,
    'DiagnosticsTap': 4.0,
    'Dism': 4.0,
    'DismApi': 4.0,
    'DismCore': 4.0,
    'DismCorePS': 4.0,
    'DismHost': 4.0,
    'DismProv': 4.0,
    'Display': 4.0,
    'DisplaySwitch': 4.0,
    'DmiProvider': 4.0,
    'DocumentPerformanceEvents': 4.0,
    'DpiScaling': 4.0,
    'DrUpdate': 4.0,
    'DsDeployRes': 4.0,
    'DscCore': 4.0,
    'DscCoreConfProv': 4.0,
    'DscCoreR': 4.0,
    'Dscpspluginwkr': 4.0,
    'DsmUserTask': 4.0,
    'Dsui': 4.0,
    'DxpTaskSync': 4.0,
    'Dxpserver': 4.0,
    'E': 4.0,
    'EAPQEC': 4.0,
    'EKAiOOPL': 4.0,
    'EKAiOSTR': 4.0,
    'EKAiOXPS': 4.0,
    'ELSCore': 4.0,
    'ERES': 4.0,
    'ETWESEProviderResources': 4.0,
    'ETWlog': 4.0,
    'EaseOfAccessDialog': 4.0,
    'EasyImgur': 4.0,
    'EditLocal': 4.0,
    'EdmGen': 4.0,
    'EncDump': 4.0,
    'English': 4.0,
    'EscMigPlugin': 4.0,
    'EssentialsConfigPluginNative': 4.0,
    'EventAggregation': 4.0,
    'EventLogMessages': 4.0,
    'EventViewer': 2.6666666666666665,
    'ExplorerFrame': 4.0,
    'ExtExport': 4.0,
    'F': 4.0,
    'FDResPub': 4.0,
```

```
    'FSDeployRes': 4.0,
    'FWPUCLNT': 4.0,
    'FX': 4.0,
    'FXSAPI': 4.0,
    'FXSDRV': 4.0,
    'FXSRES': 4.0,
    'FXSTIFF': 4.0,
    'FXSUI': 4.0,
    'FXSWZRD': 4.0,
    'Faultrep': 4.0,
    'FdDevQuery': 4.0,
    'FileAppxStreamingDataSource': 4.0,
    'FileCoAuth': 4.0,
    'FileCoAuthLib': 4.0,
    'FileSync': 4.0,
    'FileSyncApi': 4.0,
    'FileSyncClient': 4.0,
    'FileSyncConfig': 4.0,
    'FileSyncFAL': 4.0,
    'FileSyncFALWB': 4.0,
    'FileSyncSessions': 4.0,
    'FileSyncShell': 4.0,
    'FileSyncViews': 4.0,
    'FileTracker': 4.0,
    'FileTrackerUI': 4.0,
    'FirewallAPI': 4.0,
    'FirewallControlPanel': 4.0,
    'FirewallOfflineAPI': 4.0,
    'FntCache': 4.0,
    'FolderProvider': 4.0,
    'Fondue': 4.0,
    'FssmInst': 4.0,
    'FwRemoteSvr': 4.0,
    'GPOAdmin': 4.0,
    'GPOAdminCommon': 4.0,
    'GPOAdminCustom': 4.0,
    'GPRSoP': 4.0,
    'GdiPlus': 4.0,
    'GenValObj': 4.0,
    'GenericProvider': 4.0,
    'GenuineIntel': 4.0,
    'GlobCollationHost': 4.0,
    'Gui': 4.0,
    'HCSHLPR': 4.0,
    'HCSSNAP': 4.0,
    'HOSTNAME': 4.0,
    'HalExtIntcLpioDMA': 4.0,
    'HelpPane': 4.0,
    'HelpPaneProxy': 4.0,
    'HomeGroupDiagnostic': 0.0,
    'Host': 4.0,
    'HostingFilter': 4.0,
    'IBRES': 4.0,
    'IEAdvpack': 4.0,
    'IEFileInstallAI': 4.0,
    'IEShims': 4.0,
    'IHDS': 4.0,
    'IKEEXT': 4.0,
    'IMCCPHR': 4.0,
    'IMEAPIS': 4.0,
    'IMEDICAPICCPS': 4.0,
```

```
    'IMEDICTUPDATEUI': 4.0,
    'IMEFILES': 4.0,
    'IMELM': 4.0,
    'IMEPADSM': 4.0,
    'IMEPADSV': 4.0,
    'IMESEARCH': 4.0,
    'IMESEARCHDLL': 4.0,
    'IMESEARCHPS': 4.0,
    'IMETIP': 4.0,
    'IMEWDBLD': 4.0,
    'IMJKAPI': 4.0,
    'IMJPAPI': 4.0,
    'IMJPCAC': 4.0,
    'IMJPCD': 4.0,
    'IMJPCLST': 4.0,
    'IMJPCMLD': 4.0,
    'IMJPDAPI': 4.0,
    'IMJPDCT': 4.0,
    'IMJPDCTP': 4.0,
    'IMJPKDIC': 4.0,
    'IMJPLMP': 4.0,
    'IMJPPRED': 4.0,
    'IMJPSET': 4.0,
    'IMJPSKF': 4.0,
    'IMJPTIP': 4.0,
    'IMJPUEX': 4.0,
    'IMSCDICB': 4.0,
    'IMSCPROP': 4.0,
    'IMTCCAC': 4.0,
    'IMTCCFG': 4.0,
    'IMTCCORE': 4.0,
    'IMTCDIC': 4.0,
    'IMTCLNWZ': 4.0,
    'IMTCPROP': 4.0,
    'IMTCSKF': 4.0,
    'IMTCTIP': 4.0,
    'IMTCTRLN': 4.0,
    'INETRES': 4.0,
    'IPHLPAPI': 4.0,
    'IPSECSVC': 4.0,
    'IPSEventLogMsg': 4.0,
    'ISCII': 4.0,
    'ISymWrapper': 4.0,
    'IasMigPlugin': 4.0,
    'IasMigReader': 4.0,
    'IconCodecService': 4.0,
    'IdCtrls': 4.0,
    'ImSCCfg': 4.0,
    'ImSCCore': 4.0,
    'ImagingProvider': 4.0,
    'ImeBroker': 4.0,
    'ImeBrokerps': 4.0,
    'InetMgr': 4.0,
    'InfDefaultInstall': 4.0,
    'InputSwitch': 4.0,
    'Install': 4.0,
    'InstallEventRes': 4.0,
    'InstallUtil': 4.0,
    'InstallUtilLib': 4.0,
    'Installer': 4.0,
    'IntlProvider': 4.0,
```

```
    'IphlpsvcMigPlugin': 4.0,
    'JSC': 4.0,
    'JSProfilerCore': 4.0,
    'JavaScriptCollectionAgent': 4.0,
    'JpnIME': 4.0,
    'JpnKorRoaming': 4.0,
    'JpnRanker': 4.0,
    'KBDAL': 4.0,
    'KBDARME': 4.0,
    'KBDARMW': 4.0,
    'KBDAZE': 4.0,
    'KBDAZEL': 4.0,
    'KBDAZST': 4.0,
    'KBDBASH': 4.0,
    'KBDBE': 4.0,
    'KBDBENE': 4.0,
    'KBDBGPH': 4.0,
    'KBDBHC': 4.0,
    'KBDBLR': 4.0,
    'KBDBR': 4.0,
    'KBDBU': 4.0,
    'KBDBUG': 4.0,
    'KBDBULG': 4.0,
    'KBDCA': 4.0,
    'KBDCAN': 4.0,
    'KBDCHER': 4.0,
    'KBDCHERP': 4.0,
    'KBDCR': 4.0,
    'KBDCZ': 4.0,
    'KBDDA': 4.0,
    'KBDDV': 4.0,
    'KBDES': 4.0,
    'KBDEST': 4.0,
    'KBDFA': 4.0,
    'KBDFC': 4.0,
    'KBDFI': 4.0,
    'KBDFO': 4.0,
    'KBDFR': 4.0,
    'KBDFTHRK': 4.0,
    'KBDGAE': 4.0,
    'KBDGEO': 4.0,
    'KBDGKL': 4.0,
    'KBDGN': 4.0,
    'KBDGR': 4.0,
    'KBDGRLND': 4.0,
    'KBDGTHC': 4.0,
    'KBDHAU': 4.0,
    'KBDHAW': 4.0,
    'KBDHE': 4.0,
    'KBDHEB': 4.0,
    'KBDHEPT': 4.0,
    'KBDHU': 4.0,
    'KBDIBO': 4.0,
    'KBDIC': 4.0,
    'KBDINASA': 4.0,
    'KBDINBEN': 4.0,
    'KBDINDEV': 4.0,
    'KBDINEN': 4.0,
    'KBDINGUJ': 4.0,
    'KBDINHIN': 4.0,
    'KBDINKAN': 4.0,
```

```
    'KBDINMAL': 4.0,
    'KBDINMAR': 4.0,
    'KBDINORI': 4.0,
    'KBDINPUN': 4.0,
    'KBDINTAM': 4.0,
    'KBDINTEL': 4.0,
    'KBDIR': 4.0,
    'KBDIT': 4.0,
    'KBDIULAT': 4.0,
    'KBDJAV': 4.0,
    'KBDJPN': 4.0,
    'KBDKAZ': 4.0,
    'KBDKHMR': 4.0,
    'KBDKNI': 4.0,
    'KBDKOR': 4.0,
    'KBDKURD': 4.0,
    'KBDKYR': 4.0,
    'KBDLA': 4.0,
    'KBDLAO': 4.0,
    'KBDLT': 4.0,
    'KBDLV': 4.0,
    'KBDLVST': 4.0,
    'KBDMAC': 4.0,
    'KBDMACST': 4.0,
    'KBDMAORI': 4.0,
    'KBDMON': 4.0,
    'KBDMONMO': 4.0,
    'KBDMONST': 4.0,
    'KBDMYAN': 4.0,
    'KBDNE': 4.0,
    'KBDNEPR': 4.0,
    'KBDNO': 4.0,
    'KBDNSO': 4.0,
    'KBDNTL': 4.0,
    'KBDOGHAM': 4.0,
    'KBDOLCH': 4.0,
    'KBDOLDIT': 4.0,
    'KBDOSM': 4.0,
    'KBDPASH': 4.0,
    'KBDPL': 4.0,
    'KBDPO': 4.0,
    'KBDRO': 4.0,
    'KBDROPR': 4.0,
    'KBDROST': 4.0,
    'KBDRU': 4.0,
    'KBDRUM': 4.0,
    'KBDSF': 4.0,
    'KBDSG': 4.0,
    'KBDSL': 4.0,
    'KBDSMSFI': 4.0,
    'KBDSMSNO': 4.0,
    'KBDSORA': 4.0,
    'KBDSOREX': 4.0,
    'KBDSORST': 4.0,
    'KBDSP': 4.0,
    'KBDSW': 4.0,
    'KBDTAILE': 4.0,
    'KBDTAJIK': 4.0,
    'KBDTAT': 4.0,
    'KBDTIFI': 4.0,
    'KBDTIPRC': 4.0,
```

```
    'KBDTIPRD': 4.0,
    'KBDTUF': 4.0,
    'KBDTUQ': 4.0,
    'KBDTURME': 4.0,
    'KBDTZM': 4.0,
    'KBDUGHR': 4.0,
    'KBDUK': 4.0,
    'KBDUKX': 4.0,
    'KBDUR': 4.0,
    'KBDURDU': 4.0,
    'KBDUS': 4.0,
    'KBDUSA': 4.0,
    'KBDUSL': 4.0,
    'KBDUSR': 4.0,
    'KBDUSX': 4.0,
    'KBDUZB': 4.0,
    'KBDVNTC': 4.0,
    'KBDWOL': 4.0,
    'KBDYAK': 4.0,
    'KBDYBA': 4.0,
    'KBDYCC': 4.0,
    'KBDYCL': 4.0,
    'KMSVC': 4.0,
    'KdsCli': 4.0,
    'KdsSvc': 4.0,
    'KernelBase': 4.0,
    'KorHanjaDS': 4.0,
    'KorIME': 4.0,
    'KrnlProv': 4.0,
    'LBService': 4.0,
    'LXFPA': 4.0,
    'LXFPC': 4.0,
    'LXFPM': 4.0,
    'LXFPS': 4.0,
    'LXFPW': 4.0,
    'LXPJLMW': 4.0,
    'LXPTMV': 4.0,
    'LangCleanupSysprepAction': 4.0,
    'Langs': 4.0,
    'LaunchTM': 4.0,
    'Launcher': 4.0,
    'LbfoAdmin': 4.0,
    'LbfoAdminLib': 4.0,
    'LicensingUI': 4.0,
    'Local': 4.0,
    'Locator': 4.0,
    'LogMeIn': 4.0,
    'LogProvider': 4.0,
    'LoggingPlatform': 4.0,
    'LogonUI': 4.0,
    'MIGUIControls': 2.6666666666666665,
    'MMCEx': 2.6666666666666665,
    'MMCFxCommon': 2.6666666666666665,
    'MMDevAPI': 4.0,
    'MMFUtil': 4.0,
    'MPSSVC': 4.0,
    'MRINFO': 4.0,
    'MRT': 4.0,
    'MSBuild': 2.6666666666666665,
    'MSTTSEngine': 4.0,
    'MSTTSLoc': 4.0,
```

```
    'MSchedExe': 4.0,
    'MSxpsPS': 4.0,
    'MTF': 4.0,
    'MTFServer': 4.0,
    'MTFUtils': 4.0,
    'MUILanguageCleanup': 4.0,
    'MachinerySetup': 4.0,
    'Magnification': 4.0,
    'Magnify': 4.0,
    'MaintenanceUI': 4.0,
    'Marshal': 4.0,
    'McxDriv': 4.0,
    'MdRes': 4.0,
    'MdSched': 4.0,
    'MemoryAnalyzer': 4.0,
    'MemoryDiagnostic': 4.0,
    'Microsoft': 2.925,
    'MigRegDB': 4.0,
    'MirrorDrvCompat': 4.0,
    'MmcAspExt': 4.0,
    'ModataPerfCounters': 4.0,
    'ModemMigPlugin': 4.0,
    'MpSigStub': 4.0,
    'MrmCoreR': 4.0,
    'MrmIndexer': 4.0,
    'MsApoFxProxy': 4.0,
    'MsCtfMonitor': 4.0,
    'MsRdpWebAccess': 4.0,
    'MsSpellCheckingFacility': 4.0,
    'MsSpellCheckingHost': 4.0,
    'MshtmlDac': 4.0,
    'MsiCofire': 4.0,
    'MsiProvider': 4.0,
    'MuiUnattend': 4.0,
    'MultiDigiMon': 4.0,
    'Multimedia': 4.0,
    'MuxInst': 4.0,
    'N': 4.0,
    'NAPCRYPT': 4.0,
    'NAPHLPR': 4.0,
    'NAPINIT': 4.0,
    'NAPMONTR': 4.0,
    'NAPSNAP': 4.0,
    'NAPSTAT': 4.0,
    'NCProv': 4.0,
    'NETSTAT': 4.0,
    'NTFRSPRF': 4.0,
    'NXFSA': 4.0,
    'NXFSB': 4.0,
    'NapiNSP': 4.0,
    'Narrator': 4.0,
    'NativeStrings': 4.0,
    'NcaApi': 4.0,
    'NcaSvc': 4.0,
    'NcdProp': 4.0,
    'NdisImPlatform': 4.0,
    'NetAdapterCim': 4.0,
    'NetEventPacketCapture': 4.0,
    'NetEvtFwdr': 4.0,
    'NetNat': 4.0,
    'NetPeerDistCim': 4.0,
```

```
    'NetSetupAI': 4.0,
    'NetSetupApi': 4.0,
    'NetTCPIP': 4.0,
    'NetVscCoinstall': 4.0,
    'Netplwiz': 4.0,
    'NetworkDiagnosticSnapIn': 4.0,
    'NetworkStatus': 4.0,
    'Nlsdl': 4.0,
    'NzbDrone': 4.0,
    'O': 4.0,
    'OEMLicense': 4.0,
    'OKESCPU': 4.0,
    'OPEGadget': 4.0,
    'OSProvider': 4.0,
    'OnDemandConnRouteHelper': 4.0,
    'OneDrive': 4.0,
    'OneDriveSetup': 4.0,
    'OneDriveStandaloneUpdater': 4.0,
    'OpcServices': 4.0,
    'OpenWith': 4.0,
    'OskSupport': 4.0,
    'OvisFormServer': 4.0,
    'PATHPING': 4.0,
    'PCLXL': 4.0,
    'PCPKsp': 4.0,
    'PING': 4.0,
    'PJLMON': 4.0,
    'PNPXAssoc': 4.0,
    'PNPXAssocPrx': 4.0,
    'PSDSCFileDownloadManagerEvents': 4.0,
    'PSEvents': 4.0,
    'PSHED': 4.0,
    'PSModuleDiscoveryProvider': 4.0,
    'PeerDistCacheProvider': 4.0,
    'PeerDistSh': 4.0,
    'PenIMC': 4.0,
    'PerfCounter': 4.0,
    'PhotoMetadataHandler': 4.0,
    'PickerHost': 4.0,
    'PkgMgr': 4.0,
    'PlaMig': 4.0,
    'PlaySndSrv': 4.0,
    'PlayToStatusProvider': 4.0,
    'PnPUnattend': 4.0,
    'PnPutil': 4.0,
    'PolicMan': 4.0,
    'Policy': 4.0,
    'PortQryUI': 4.0,
    'PortableDeviceSyncProvider': 4.0,
    'PowerWmiProvider': 0.0,
    'PresentationBuildTasks': 2.6666666666666665,
    'PresentationCore': 2.6666666666666665,
    'PresentationFramework': 2.1333333333333333,
    'PresentationHost': 4.0,
    'PresentationHostProxy': 4.0,
    'PresentationUI': 2.6666666666666665,
    'PrintAdvancedInstaller': 4.0,
    'PrintConfig': 4.0,
    'PrintDialogHost': 4.0,
    'PrintDialogs': 4.0,
    'PrintIsolationHost': 4.0,
```

```
    'PrintIsolationProxy': 4.0,
    'PrintManagementProvider': 4.0,
    'Pritunl': 4.0,
    'Provider': 4.0,
    'PwdSSP': 4.0,
    'Q': 4.0,
    'QAGENT': 4.0,
    'QAGENTRT': 4.0,
    'QCLIPROV': 4.0,
    'QSHVHOST': 4.0,
    'QSVRMGMT': 4.0,
    'QUTIL': 4.0,
    'Query': 4.0,
    'RADCUI': 4.0,
    'RAMgmtPSProvider': 4.0,
    'RAMgmtUI': 4.0,
    'RASMM': 4.0,
    'RAServerPSProvider': 4.0,
    'RC': 4.0,
    'RDSAppXHelper': 4.0,
    'RDSPnf': 4.0,
    'RDWebAI': 4.0,
    'RDWebServiceAsp': 4.0,
    'RES': 4.0,
    'RMActivate': 4.0,
    'ROUTE': 4.0,
    'RSSme': 4.0,
    'RTWorkQ': 4.0,
    'RacEngn': 4.0,
    'RacWmiProv': 4.0,
    'RasClusterRes': 4.0,
    'RasMigPlugin': 4.0,
    'RdmsInst': 4.0,
    'RdpSa': 4.0,
    'RdpSaProxy': 4.0,
    'RdpSaPs': 4.0,
    'RdpSaUacHelper': 4.0,
    'ReAgent': 4.0,
    'ReAgentc': 4.0,
    'ReInfo': 4.0,
    'ReachFramework': 2.6666666666666665,
    'RegAsm': 4.0,
    'RegCtrl': 4.0,
    'RegSvcs': 4.0,
    'Regasm': 4.0,
    'RegisterIEPKeysAI': 4.0,
    'RelPost': 4.0,
    'RemoteAccessDbVerification': 4.0,
    'RemoveDeviceContextHandler': 4.0,
    'RemoveDeviceElevated': 4.0,
    'ResourceDll': 4.0,
    'Resources': 4.0,
    'RmClient': 4.0,
    'RoamingSecurity': 4.0,
    'Robocopy': 4.0,
    'RpcEpMap': 4.0,
    'RpcPing': 4.0,
    'RpcProxy': 4.0,
    'RpcProxyMigrationPlugin': 4.0,
    'RpcRtRemote': 4.0,
    'RstrtMgr': 4.0,
```

```
    'RunLegacyCPLElevated': 4.0,
    'RuntimeBroker': 4.0,
    'S': 4.0,
    'SCGMigPlugin': 4.0,
    'SCW': 4.0,
    'SCWViewer': 4.0,
    'SCardDlg': 4.0,
    'SCardSvr': 4.0,
    'SDClient': 4.0,
    'SETUP': 4.0,
    'SFTPMSI': 4.0,
    'SHCore': 4.0,
    'SMBHelperClass': 4.0,
    'SMDiagnostics': 2.0,
    'SMEF': 4.0,
    'SMRemoting': 4.0,
    'SMSvcHost': 2.6666666666666665,
    'SMTPCons': 4.0,
    'SMdiagnostics': 4.0,
    'SOS': 4.0,
    'SPInf': 4.0,
    'SQLOS': 4.0,
    'SQLSCM': 4.0,
    'SRH': 4.0,
    'SSShim': 4.0,
    'SamepageSetup': 4.0,
    'SbsNclPerf': 4.0,
    'ScDeviceEnum': 4.0,
    'SceneCefBrowser': 4.0,
    'ScwAuditExt': 4.0,
    'ScwFirewallExt': 4.0,
    'ScwRegistryExt': 4.0,
    'ScwSceExt': 4.0,
    'ScwServiceExt': 4.0,
    'Scylla': 4.0,
    'SeVA': 4.0,
    'SearchFolder': 4.0,
    'SecEdit': 4.0,
    'SecHC': 4.0,
    'SecurityAuditPoliciesSnapIn': 2.6666666666666665,
    'Sens': 4.0,
    'SensApi': 4.0,
    'ServDeps': 4.0,
    'ServerCeipOptin': 4.0,
    'ServerCeipOptinGui': 4.0,
    'ServerManager': 4.0,
    'ServerManagerLauncher': 4.0,
    'ServerTelemetryConfig': 4.0,
    'ServerWerOptin': 4.0,
    'ServerWerOptinGui': 4.0,
    'ServiceModelEvents': 4.0,
    'ServiceModelInstallRC': 4.0,
    'ServiceModelPerformanceCounters': 4.0,
    'ServiceModelReg': 4.0,
    'ServiceModelRegMigPlugin': 4.0,
    'ServiceModelRegUI': 4.0,
    'ServiceMonikerSupport': 4.0,
    'SessEnv': 4.0,
    'SetIEInstalledDateAI': 4.0,
    'SetNetworkLocation': 4.0,
    'SetProxyCredential': 4.0,
```

```
    'Setup': 4.0,
    'ShellSetup': 4.0,
    'SmartCardSimulator': 4.0,
    'SmartScreenSettings': 4.0,
    'SmartcardCredentialProvider': 4.0,
    'SmiProvider': 4.0,
    'SmisConfigProvRes': 4.0,
    'SndVol': 4.0,
    'SndVolSSO': 4.0,
    'SpaceAgent': 4.0,
    'SpeechUXRes': 4.0,
    'SppExtComObj': 4.0,
    'SppMig': 4.0,
    'Sql': 4.0,
    'SqlAccess': 4.0,
    'SqlDK': 4.0,
    'SqlDumper': 4.0,
    'SqlTsEs': 4.0,
    'SrpUxNativeSnapIn': 4.0,
    'SrpUxSnapIn': 2.6666666666666665,
    'SrvMgrInst': 4.0,
    'StarCraft': 4.0,
    'StarEdit': 4.0,
    'StaticDictDS': 4.0,
    'SteamSetup': 4.0,
    'StorMigPlugin': 4.0,
    'StorageContextHandler': 4.0,
    'StorageServiceRes': 4.0,
    'Storprop': 4.0,
    'StructuredQuery': 4.0,
    'SxsMigPlugin': 4.0,
    'SyncHost': 4.0,
    'SyncHostps': 4.0,
    'SyncInfrastructure': 4.0,
    'SyncInfrastructureps': 4.0,
    'SyncShareRes': 4.0,
    'Syncreg': 4.0,
    'SysFxUI': 4.0,
    'System': 2.8372093023255816,
    'SystemCore': 2.0,
    'SystemData': 2.0,
    'SystemDrawing': 2.0,
    'SystemEventsBrokerServer': 4.0,
    'SystemPropertiesAdvanced': 4.0,
    'SystemPropertiesComputerName': 4.0,
    'SystemPropertiesDataExecutionPrevention': 4.0,
    'SystemPropertiesHardware': 4.0,
    'SystemPropertiesPerformance': 4.0,
    'SystemPropertiesProtection': 4.0,
    'SystemPropertiesRemote': 4.0,
    'SystemSurvey': 4.0,
    'SystemXml': 2.0,
    'SystemXmlLinq': 2.0,
    'T': 4.0,
    'TCPSVCS': 4.0,
    'TLBREF': 4.0,
    'TRACERT': 4.0,
    'TRLParserCOMInterface': 2.0,
    'TSChannel': 4.0,
    'TSErrRedir': 4.0,
    'TSPSCmdlets': 4.0,
```

```
    'TSPSDataAccess': 4.0,
    'TSPSEngine': 4.0,
    'TSPSProvider': 4.0,
    'TSPortalWebPart': 4.0,
    'TSTheme': 4.0,
    'TSWbPrxy': 4.0,
    'TSWorkspace': 4.0,
    'TSpkg': 4.0,
    'TTY': 4.0,
    'TTYRES': 4.0,
    'TTYUI': 4.0,
    'Tabbtn': 4.0,
    'TabbtnEx': 4.0,
    'TableTextService': 4.0,
    'TableTextServiceMig': 4.0,
    'TapiMigPlugin': 4.0,
    'TapiSysprep': 4.0,
    'TapiUnattend': 4.0,
    'TaskSchdPS': 4.0,
    'TaskScheduler': 2.6666666666666665,
    'Taskmgr': 4.0,
    'TcpipSetup': 4.0,
    'ThumbnailExtractionHost': 4.0,
    'TiFileFetcher': 4.0,
    'TiWorker': 4.0,
    'TieringEngineProxy': 4.0,
    'TieringEngineService': 4.0,
    'TimeDateMUICallback': 4.0,
    'Timeline': 4.0,
    'TitanEngine': 4.0,
    'TlsBrand': 4.0,
    'Tools': 4.0,
    'Tpm': 4.0,
    'TpmInit': 4.0,
    'TpmTasks': 4.0,
    'TransformationRulesParser': 4.0,
    'TransmogProvider': 4.0,
    'Tresorit': 4.0,
    'TrustedInstaller': 4.0,
    'TsPnPRdrCoInstaller': 4.0,
    'TsPsUtil': 4.0,
    'TsUsbGDCoInstaller': 4.0,
    'TsUsbRedirectionGroupPolicyExtension': 4.0,
    'TtlsAuth': 4.0,
    'TtlsCfg': 4.0,
    'TtlsExt': 4.0,
    'U': 4.0,
    'UI': 4.0,
    'UIAnimation': 4.0,
    'UIAutomationClient': 2.6666666666666665,
    'UIAutomationClientsideProviders': 2.6666666666666665,
    'UIAutomationCore': 4.0,
    'UIAutomationCoreRes': 4.0,
    'UIAutomationProvider': 2.6666666666666665,
    'UIAutomationTypes': 2.6666666666666665,
    'UIImg': 4.0,
    'UIRes': 4.0,
    'UIRibbon': 4.0,
    'UIRibbonRes': 4.0,
    'UNIDRV': 4.0,
    'UNIDRVUI': 4.0,
```

```
    'UNIRES': 4.0,
    'URES': 4.0,
    'UXInit': 4.0,
    'UnattendProvider': 4.0,
    'Uninstall': 4.0,
    'Uninstaller': 4.0,
    'UserAccountBroker': 4.0,
    'UserAccountControlSettings': 4.0,
    'UserLanguageProfileCallback': 4.0,
    'UserLanguagesCpl': 4.0,
    'Utilman': 4.0,
    'Utils': 4.0,
    'VAN': 4.0,
    'VBoxControl': 4.0,
    'VBoxDisp': 4.0,
    'VBoxDrvInst': 4.0,
    'VBoxHook': 4.0,
    'VBoxMRXNP': 4.0,
    'VBoxOGL': 4.0,
    'VBoxOGLarrayspu': 4.0,
    'VBoxOGLcrutil': 4.0,
    'VBoxOGLerrorspu': 4.0,
    'VBoxOGLfeedbackspu': 4.0,
    'VBoxOGLpackspu': 4.0,
    'VBoxOGLpassthroughspu': 4.0,
    'VBoxService': 4.0,
    'VBoxTray': 4.0,
    'VBoxWHQLFake': 4.0,
    'VGX': 4.0,
    'VMBusVideoD': 4.0,
    'VSSUI': 4.0,
    'VSSUIRUN': 4.0,
    'VSSVC': 4.0,
    'Vault': 4.0,
    'VaultCmd': 4.0,
    'VaultRoaming': 4.0,
    'VhdProvider': 4.0,
    'VirtualSmartcardReader': 4.0,
    'VmApplicationHealthMonitorProxy': 4.0,
    'VmHostAI': 4.0,
    'VmbusCoinstaller': 4.0,
    'VmdCoinstall': 4.0,
    'VscMgrPS': 4.0,
    'W': 4.0,
    'WABSyncProvider': 4.0,
    'WINDOWS': 4.0,
    'WINSEVNT': 4.0,
    'WINSRPC': 4.0,
    'WLanHC': 4.0,
    'WMALFXGFXDSP': 4.0,
    'WMIADAP': 4.0,
    'WMIC': 4.0,
    'WMICOOKR': 4.0,
    'WMIMigrationPlugin': 4.0,
    'WMIPICMP': 4.0,
    'WMIPIPRT': 4.0,
    'WMIPJOBJ': 4.0,
    'WMIPSESS': 4.0,
    'WMIsvc': 4.0,
    'WMPhoto': 4.0,
    'WMSPDMOD': 4.0,
```

```
 'WSClient': 4.0,
 'WSDApi': 4.0,
 'WSDMon': 4.0,
 'WSDPrintProxy': 4.0,
 'WSDScDrv': 4.0,
 'WSDScanProxy': 4.0,
 'WSEDeployRes': 4.0,
 'WSHTCPIP': 4.0,
 'WSManHTTPConfig': 4.0,
 'WSManMigrationPlugin': 4.0,
 'WSMigPlugin': 4.0,
 'WSSG': 4.0,
 'WSService': 4.0,
 'WSShared': 4.0,
 'WSSync': 4.0,
 'WUDFCoinstaller': 4.0,
 'WUDFHost': 4.0,
 'WUDFPlatform': 4.0,
 'WUDFSvc': 4.0,
 'WUDFUsbccidDriver': 4.0,
 'WUDFx': 4.0,
 'WUSettingsProvider': 4.0,
 'WWAHost': 4.0,
 'WallpaperHost': 4.0,
 'WcsPlugInService': 4.0,
 'WdacWmiProv': 4.0,
 ...}
```

In [33]:

```
# Impute ages based on titles
idx_nan_age = df.loc[np.isnan(df['e_cparhdr'])].index
df.loc[idx_nan_age,'e_cparhdr'].loc[idx_nan_age] = df['Title'].loc[idx_nan_age].map(map
_means)
df.head()
```

Out[33]:

	Name	e_cparhdr	e_maxalloc	e_ss	e_sp	e_csum
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	65535	0	184	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	65535	0	184	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	65535	0	184	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	65535	0	184	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	65535	0	184	0

5 rows × 75 columns

In [34]:

```
# Identify imputed data
df['Imputed'] = 0
df.at[idx_nan_age.values, 'Imputed'] = 1
df.head()
```

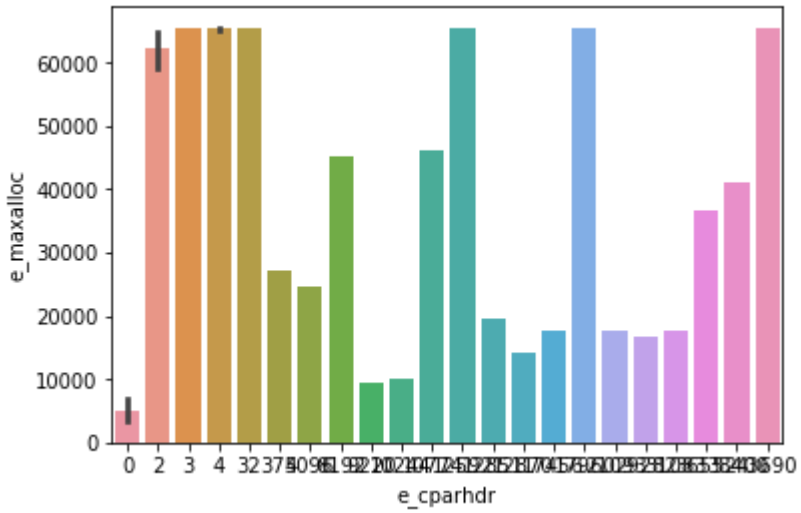
Out[34]:

	Name	e_cparhdr	e_maxalloc	e_ss	e_sp	e_csum
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	65535	0	184	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	65535	0	184	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	65535	0	184	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	65535	0	184	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	65535	0	184	0

5 rows × 76 columns

In [35]:

```
# Plot
sns.barplot(df['e_cparhdr'],df['e_maxalloc']);
```



In [36]:

```
# Count how many people have each of the titles
df.groupby(['e_ss'])['e_sp'].count()
```

Out[36]:

```
e_ss
0        19596
2            3
7            1
11           1
62           1
9216         1
10496        1
11948        1
12752        2
13378        1
16950        1
55303        1
61436        1
Name: e_sp, dtype: int64
```

In [37]:

```
titles_dict = {'Anuradha': 'Other',
               'Keran': 'Other',
               'Keya': 'Other',
               'Koyeli': 'Other',
               'Adrija': 'Other',
               'FATIMA': 'Other',
               'Mohit dhiman': 'Other',
               'Mohit Sharma': 'Other',
               'Ayush': 'Other',
               'Deepanjali': 'Mrs',
               'Shekhar': 'Miss',
               'Sanand': 'Miss',
               'Vaibhav': 'Mr',
               'Lavisha': 'Mrs',
               'Vikas': 'Miss',
               'Akhil': 'Master',
               'RS BAWA': 'Other'}
```

In [38]:

```
# Group titles
df['Title'] = df['Title'].map(titles_dict)
df['Title'].head()
```

Out[38]:

```
0    NaN
1    NaN
2    NaN
3    NaN
4    NaN
Name: Title, dtype: object
```

In [39]:

```
# Transform into categorical
df['Title'] = pd.Categorical(df['Title'])
df.dtypes
```

Out[39]:

```
Name                                object
e_cparhdr                            int64
e_maxalloc                           int64
e_ss                                 int64
e_sp                                 int64
e_csum                               int64
e_ip                                 int64
e_cs                                 int64
e_lfarlc                             int64
e_ovno                               int64
e_oemid                              int64
e_oeminfo                            int64
e_lfanew                             int64
Machine                              int64
NumberOfSections                     int64
TimeDateStamp                        int64
PointerToSymbolTable                 int64
NumberOfSymbols                      int64
SizeOfOptionalHeader                 int64
Characteristics                      int64
Magic                                int64
MajorLinkerVersion                   int64
MinorLinkerVersion                   int64
SizeOfCode                           int64
SizeOfInitializedData                int64
SizeOfUninitializedData              int64
AddressOfEntryPoint                  int64
BaseOfCode                           int64
ImageBase                            int64
SectionAlignment                     int64
                                    ...
LoaderFlags                          int64
NumberOfRvaAndSizes                  int64
Malware                              int64
SuspiciousImportFunctions            int64
SuspiciousNameSection                int64
SectionsLength                       int64
SectionMinEntropy                  float64
SectionMaxEntropy                    int64
SectionMinRawsize                    int64
SectionMaxRawsize                    int64
SectionMinVirtualsize                int64
SectionMaxVirtualsize                int64
SectionMaxPhysical                   int64
SectionMinPhysical                   int64
SectionMaxVirtual                    int64
SectionMinVirtual                    int64
SectionMaxPointerData                int64
SectionMinPointerData                int64
SectionMaxChar                       int64
SectionMainChar                      int64
DirectoryEntryImport                 int64
DirectoryEntryImportSize             int64
DirectoryEntryExport                 int64
ImageDirectoryEntryExport            int64
ImageDirectoryEntryImport            int64
ImageDirectoryEntryResource          int64
ImageDirectoryEntryException         int64
ImageDirectoryEntrySecurity          int64
```

```
Title                          category
Imputed                           int64
Length: 76, dtype: object
```

In [42]:

```
# Plot
sns.barplot(x='Title', y='e_ss', data=df);
```

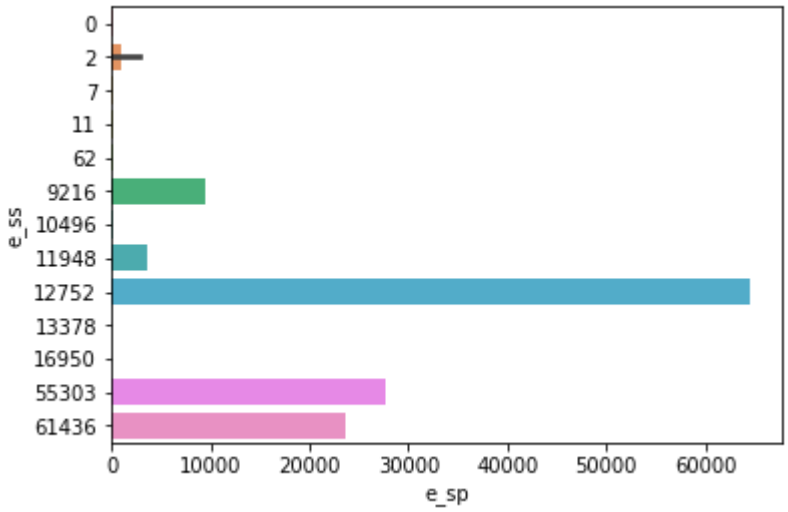
```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call las
t)
<ipython-input-42-d4d4dd602ac1> in <module>
      1 # Plot
----> 2 sns.barplot(x='Title', y='e_ss', data=df);
      3

~\Anaconda3\lib\site-packages\seaborn\categorical.py in barplot(x, y, hue,
data, order, hue_order, estimator, ci, n_boot, units, orient, color, palet
te, saturation, errcolor, errwidth, capsize, dodge, ax, **kwargs)
   3147                          estimator, ci, n_boot, units,
   3148                          orient, color, palette, saturation,
-> 3149                          errcolor, errwidth, capsize, dodge)
   3150
   3151     if ax is None:

~\Anaconda3\lib\site-packages\seaborn\categorical.py in __init__(self, x,
 y, hue, data, order, hue_order, estimator, ci, n_boot, units, orient, col
or, palette, saturation, errcolor, errwidth, capsize, dodge)
   1606         self.establish_variables(x, y, hue, data, orient,
   1607                                  order, hue_order, units)
-> 1608         self.establish_colors(color, palette, saturation)
   1609         self.estimate_statistic(estimator, ci, n_boot)
   1610

~\Anaconda3\lib\site-packages\seaborn\categorical.py in establish_colors(s
elf, color, palette, saturation)
    313         # Determine the gray color to use for the lines framing th
e plot
    314         light_vals = [colorsys.rgb_to_hls(*c)[1] for c in rgb_colo
rs]
--> 315         lum = min(light_vals) * .6
    316         gray = mpl.colors.rgb2hex((lum, lum, lum))
    317

ValueError: min() arg is an empty sequence
```

In [44]:

```
# Transform into categorical
df['e_sp'] = pd.Categorical(df['e_sp'])
```

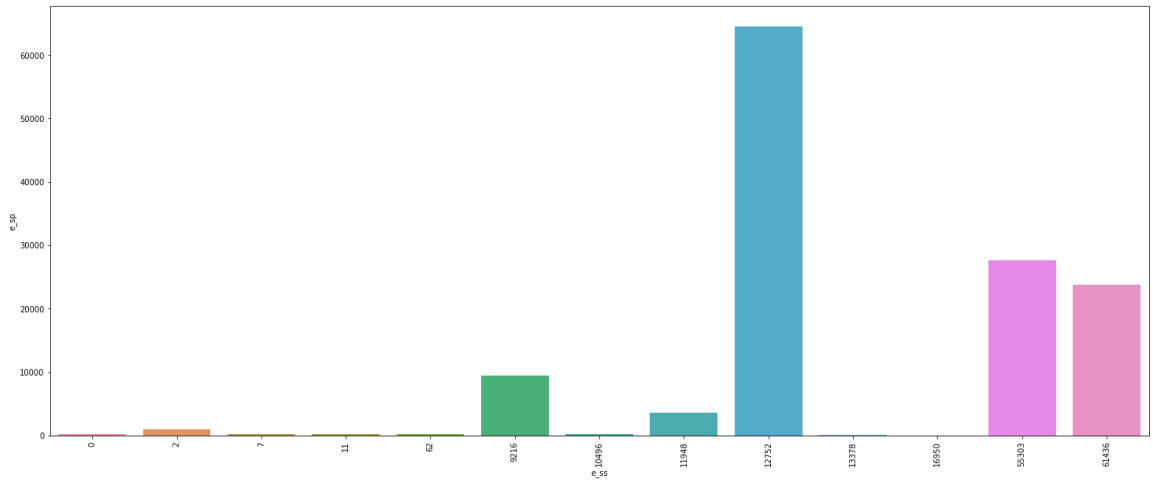
In [45]:

```
# Plot
sns.barplot(df['e_sp'],df['e_ss']);
```



In [46]:

```
# Plot
plt.figure(figsize=(25,10))
sns.barplot(df['e_ss'],df['e_sp'], ci=None)
plt.xticks(rotation=90);
```



In [47]:

```
# Plot
'''
Probably, there is an easier way to do this plot. I had a problem using
plt.axvspan because the xmin and xmax values weren't
being plotted correctly. For example, I would define xmax = 12 and only
the area between 0 and 7 would be filled. This was happening because my
X-axis don't follow a regular (0, 1, ..., n) sequence. After some trial
and error, I noticed that xmin and xmax refer to the number of elements in
the X-axis coordinate that should be filled. Accordingly, I defined two
variables, x_limit_1 and x_limit_2, that count the number of elements that
should be filled in each interval. Sounds confusing? To me too.
'''
limit_1 = 12
limit_2 = 50

x_limit_1 = np.size(df[df['e_ss'] < limit_1]['e_ss'].unique())
x_limit_2 = np.size(df[df['e_ss'] < limit_2]['e_ss'].unique())

plt.figure(figsize=(25,10))
sns.barplot(df['e_ss'],df['e_sp'], ci=None)

plt.axvspan(-1, x_limit_1, alpha=0.25, color='green')
plt.axvspan(x_limit_1, x_limit_2, alpha=0.25, color='red')
plt.axvspan(x_limit_2, 100, alpha=0.25, color='yellow')

plt.xticks(rotation=90);
```

```
---------------------------------------------------------------------------
-
TypeError                                 Traceback (most recent call las
t)
<ipython-input-47-1b81fb89320d> in <module>
     14 limit_2 = 50
     15
---> 16 x_limit_1 = np.size(df[df['e_ss'] < limit_1]['e_ss'].unique())
     17 x_limit_2 = np.size(df[df['e_ss'] < limit_2]['e_ss'].unique())
     18

~\Anaconda3\lib\site-packages\pandas\core\ops.py in wrapper(self, other, a
xis)
   1194             # Dispatch to Categorical implementation; pd.Categoric
alIndex
   1195             # behavior is non-canonical GH#19513
-> 1196             res_values = dispatch_to_index_op(op, self, other, pd.
Categorical)
   1197             return self._constructor(res_values, index=self.index,
   1198                                      name=res_name)

~\Anaconda3\lib\site-packages\pandas\core\ops.py in dispatch_to_index_op(o
p, left, right, index_class)
   1099         left_idx = left_idx._shallow_copy(freq=None)
   1100     try:
-> 1101         result = op(left_idx, right)
   1102     except NullFrequencyError:
   1103         # DatetimeIndex and TimedeltaIndex with freq == None raise
ValueError

~\Anaconda3\lib\site-packages\pandas\core\arrays\categorical.py in f(self,
other)
     73         if not self.ordered:
     74             if op in ['__lt__', '__gt__', '__le__', '__ge__']:
---> 75                 raise TypeError("Unordered Categoricals can only c
ompare "
     76                                 "equality or not")
     77         if isinstance(other, Categorical):

TypeError: Unordered Categoricals can only compare equality or not
```

In [48]:

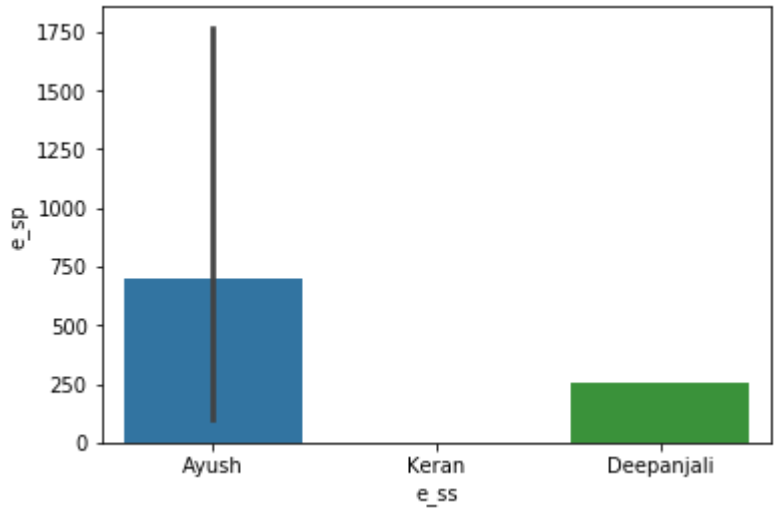
```
# Bin data
df['e_ss'] = pd.cut(df['e_ss'], bins=[0, 12, 50, 200], labels=['Ayush','Keran','Deepanj
ali'])
df['e_ss'].head()
```

Out[48]:

```
0    NaN
1    NaN
2    NaN
3    NaN
4    NaN
Name: e_ss, dtype: category
Categories (3, object): [Ayush < Keran < Deepanjali]
```

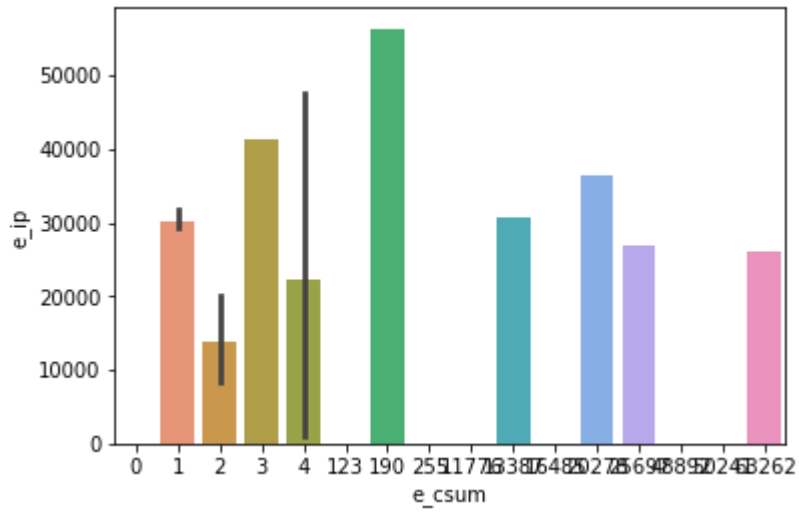
In [49]:

```
# Plot
sns.barplot(df['e_ss'], df['e_sp']);
```



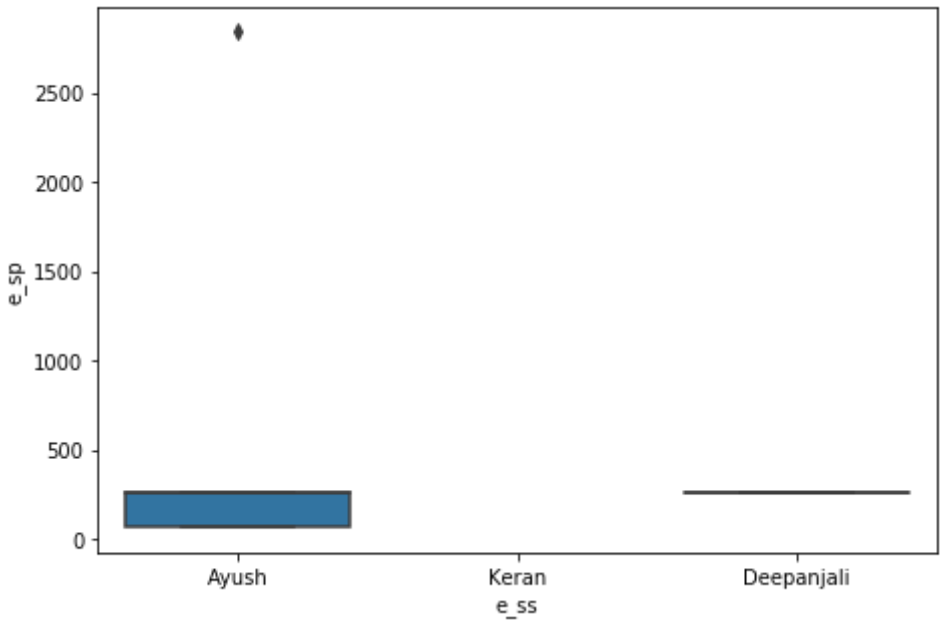
In [51]:

```
# Plot
sns.barplot(df['e_csum'], df['e_ip']);
```



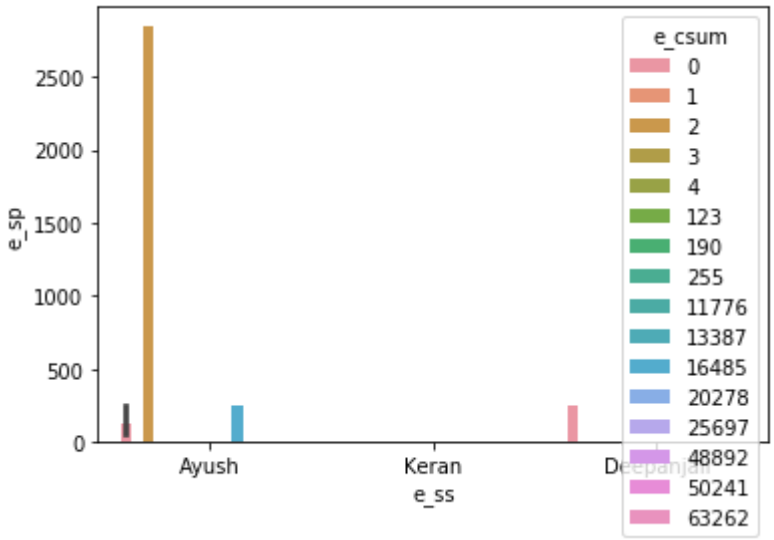
In [52]:

```
# Plot
plt.figure(figsize=(7.5,5))
sns.boxplot(df['e_ss'], df['e_sp']);
```



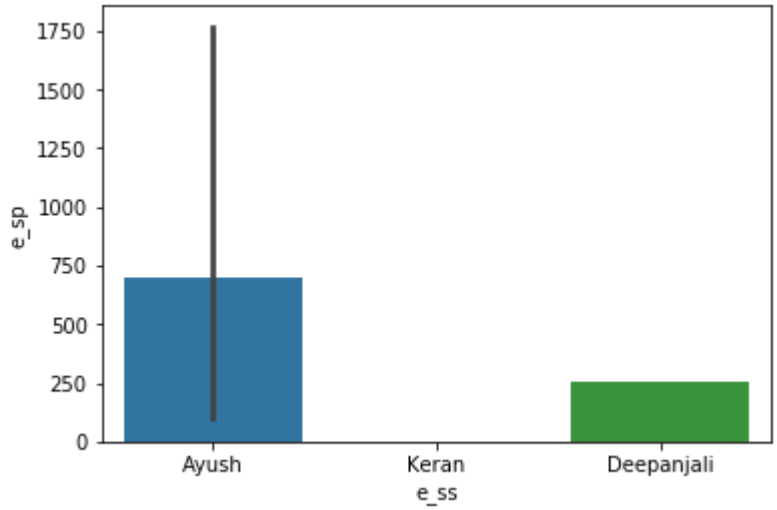
In [53]:

```
# Plot
sns.barplot(df['e_ss'], df['e_sp'], df['e_csum']);
```



In [54]:

```
# Plot
sns.barplot(df['e_ss'], df['e_sp']);
```



In [55]:

```
# Compare with other variables
df.groupby(['e_ss']).mean()
```

Out[55]:

e_ss	e_cparhdr	e_maxalloc	e_sp	e_csum	e_ip	e_cs	e_lfarlc	e_ovno	e_oemid
Ayush	821.6	57374.2	696.2	3297.4	6152.6	0.6	528.4	0.2	14868.0
Keran	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Deepanjali	32.0	65535.0	256.0	0.0	0.0	0.0	64.0	0.0	0.0

3 rows × 73 columns

In [56]:

```
# Relationship with age
df.groupby(['e_ss','e_sp'])['e_csum'].count()
```

Out[56]:

```
e_ss        e_sp
Ayush       64      2
            256     2
            2841    1
Deepanjali  256     1
Name: e_csum, dtype: int64
```

In [57]:

```
# Relationship with sex
df.groupby(['e_ss','e_sp'])['e_cs'].count()
```

Out[57]:

```
e_ss        e_sp
Ayush       64      2
            256     2
            2841    1
Deepanjali  256     1
Name: e_cs, dtype: int64
```

In [58]:

```
# Overview
df.head()
```

Out[58]:

	Name	e_cparhdr	e_maxalloc	e_ss	e_sp	e_csum
0	VirusShare_a878ba26000edaac5c98eff4432723b3	4	65535	NaN	184	0
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	4	65535	NaN	184	0
2	VirusShare_ef84cdeba22be72a69b198213dada81a	4	65535	NaN	184	0
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	4	65535	NaN	184	0
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	4	65535	NaN	184	0

5 rows × 76 columns

In [59]:

```
# Drop feature
df.drop('e_cparhdr', axis=1, inplace=True)
```

In [60]:

```
# Check features type
df.dtypes
```

Out[60]:

```
Name                              object
e_maxalloc                         int64
e_ss                            category
e_sp                               int64
e_csum                             int64
e_ip                               int64
e_cs                               int64
e_lfarlc                           int64
e_ovno                             int64
e_oemid                            int64
e_oeminfo                          int64
e_lfanew                           int64
Machine                            int64
NumberOfSections                   int64
TimeDateStamp                      int64
PointerToSymbolTable               int64
NumberOfSymbols                    int64
SizeOfOptionalHeader               int64
Characteristics                    int64
Magic                              int64
MajorLinkerVersion                 int64
MinorLinkerVersion                 int64
SizeOfCode                         int64
SizeOfInitializedData              int64
SizeOfUninitializedData            int64
AddressOfEntryPoint                int64
BaseOfCode                         int64
ImageBase                          int64
SectionAlignment                   int64
FileAlignment                      int64
                                     ...
LoaderFlags                        int64
NumberOfRvaAndSizes                int64
Malware                            int64
SuspiciousImportFunctions          int64
SuspiciousNameSection              int64
SectionsLength                     int64
SectionMinEntropy                float64
SectionMaxEntropy                  int64
SectionMinRawsize                  int64
SectionMaxRawsize                  int64
SectionMinVirtualsize              int64
SectionMaxVirtualsize              int64
SectionMaxPhysical                 int64
SectionMinPhysical                 int64
SectionMaxVirtual                  int64
SectionMinVirtual                  int64
SectionMaxPointerData              int64
SectionMinPointerData              int64
SectionMaxChar                     int64
SectionMainChar                    int64
DirectoryEntryImport               int64
DirectoryEntryImportSize           int64
DirectoryEntryExport               int64
ImageDirectoryEntryExport          int64
ImageDirectoryEntryImport          int64
ImageDirectoryEntryResource        int64
ImageDirectoryEntryException       int64
ImageDirectoryEntrySecurity        int64
```

```
Title                          category
Imputed                           int64
Length: 75, dtype: object
```

In [61]:

```
# Transform object into categorical
df['e_ss'] = pd.Categorical(df['e_sp'])
df['e_csum'] = pd.Categorical(df['e_ip'])
df.dtypes
```

Out[61]:

```
Name                                object
e_maxalloc                           int64
e_ss                              category
e_sp                                 int64
e_csum                            category
e_ip                                 int64
e_cs                                 int64
e_lfarlc                             int64
e_ovno                               int64
e_oemid                              int64
e_oeminfo                            int64
e_lfanew                             int64
Machine                              int64
NumberOfSections                     int64
TimeDateStamp                        int64
PointerToSymbolTable                 int64
NumberOfSymbols                      int64
SizeOfOptionalHeader                 int64
Characteristics                      int64
Magic                                int64
MajorLinkerVersion                   int64
MinorLinkerVersion                   int64
SizeOfCode                           int64
SizeOfInitializedData                int64
SizeOfUninitializedData              int64
AddressOfEntryPoint                  int64
BaseOfCode                           int64
ImageBase                            int64
SectionAlignment                     int64
FileAlignment                        int64
                                    ...
LoaderFlags                          int64
NumberOfRvaAndSizes                  int64
Malware                              int64
SuspiciousImportFunctions            int64
SuspiciousNameSection                int64
SectionsLength                       int64
SectionMinEntropy                  float64
SectionMaxEntropy                    int64
SectionMinRawsize                    int64
SectionMaxRawsize                    int64
SectionMinVirtualsize                int64
SectionMaxVirtualsize                int64
SectionMaxPhysical                   int64
SectionMinPhysical                   int64
SectionMaxVirtual                    int64
SectionMinVirtual                    int64
SectionMaxPointerData                int64
SectionMinPointerData                int64
SectionMaxChar                       int64
SectionMainChar                      int64
DirectoryEntryImport                 int64
DirectoryEntryImportSize             int64
DirectoryEntryExport                 int64
ImageDirectoryEntryExport            int64
ImageDirectoryEntryImport            int64
ImageDirectoryEntryResource          int64
ImageDirectoryEntryException         int64
ImageDirectoryEntrySecurity          int64
```

```
Title                          category
Imputed                           int64
Length: 75, dtype: object
```

In [62]:

```
# Transform categorical features into dummy variables
df = pd.get_dummies(df, drop_first=1)
df.head()
```

Out[62]:

	e_maxalloc	e_sp	e_ip	e_cs	e_lfarlc	e_ovno	e_oemid	e_oeminfo	e_lfanew	Machine
0	65535	184	0	0	64	0	0	0	248	34404
1	65535	184	0	0	64	0	0	0	240	332
2	65535	184	0	0	64	0	0	0	256	332
3	65535	184	0	0	64	0	0	0	128	332
4	65535	184	0	0	64	0	0	0	128	332

5 rows × 19733 columns

In [63]:

```
from sklearn.model_selection import train_test_split

X = df[df.loc[:, df.columns != 'e_cs'].columns]
y = df['e_cs']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.2, random_state=0)
```

In [82]:

```
from scipy.stats import boxcox

X_train_transformed = X_train.copy()
X_train_transformed['e_ip'] = boxcox(X_train_transformed['e_ip'] + 1)[0]
X_test_transformed = X_test.copy()
X_test_transformed['e_ip'] = boxcox(X_test_transformed['e_ip'] + 1)[0]
```

In [83]:

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X_train_transformed_scaled = scaler.fit_transform(X_train_transformed)
X_test_transformed_scaled = scaler.transform(X_test_transformed)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\preprocessing\data.p
y:323: DataConversionWarning: Data with input dtype uint8, int64, float64
were all converted to float64 by MinMaxScaler.
  return self.partial_fit(X, y)
```

In [84]:

```
from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(degree=2).fit(X_train_transformed)
X_train_poly = poly.transform(X_train_transformed_scaled)
X_test_poly = poly.transform(X_test_transformed_scaled)
```

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
<ipython-input-84-c5ed3d20c625> in <module>
      2
      3 poly = PolynomialFeatures(degree=2).fit(X_train_transformed)
----> 4 X_train_poly = poly.transform(X_train_transformed_scaled)
      5 X_test_poly = poly.transform(X_test_transformed_scaled)

~\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py in transform(self, X)
   1484             XP = sparse.hstack(columns, dtype=X.dtype).tocsc()
   1485         else:
-> 1486             XP = np.empty((n_samples, self.n_output_features_), dtype=X.dtype)
   1487             for i, comb in enumerate(combinations):
   1488                 XP[:, i] = X[:, comb].prod(1)

MemoryError:
```

In [68]:

```
# Debug
print(poly.get_feature_names())
```

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
<ipython-input-68-afb03e81658e> in <module>
      1 # Debug
----> 2 print(poly.get_feature_names())

~\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py in get_feature_names(self, input_features)
   1409
   1410         """
-> 1411         powers = self.powers_
   1412         if input_features is None:
   1413             input_features = ['x%d' % i for i in range(powers.shape[1])]

~\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py in powers_(self)
   1392                                             self.include_bias)
   1393         return np.vstack([np.bincount(c, minlength=self.n_input_features_)
-> 1394                           for c in combinations])
   1395
   1396     def get_feature_names(self, input_features=None):

~\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py in <listcomp>(.0)
   1392                                             self.include_bias)
   1393         return np.vstack([np.bincount(c, minlength=self.n_input_features_)
-> 1394                           for c in combinations])
   1395
   1396     def get_feature_names(self, input_features=None):

MemoryError:
```

In [77]:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

## Get score using original model
logreg = LogisticRegression(C=1)
logreg.fit(X_train, y_train)
scores = cross_val_score(logreg, X_train, y_train, cv=10)
print('CV accuracy (original): %.3f +/- %.3f' % (np.mean(scores), np.std(scores)))
highest_score = np.mean(scores)

## Get score using models with feature selection
for i in range(1, X_train_poly.shape[1]+1, 1):
    # Select i features
    select = SelectKBest(score_func=chi2, k=i)
    select.fit(X_train_poly, y_train)
    X_train_poly_selected = select.transform(X_train_poly)

    # Model with i features selected
    logreg.fit(X_train_poly_selected, y_train)
    scores = cross_val_score(logreg, X_train_poly_selected, y_train, cv=10)
    print('CV accuracy (number of features = %i): %.3f +/- %.3f' % (i,
                                                                    np.mean(scores),
                                                                    np.std(scores)))

    # Save results if best score
    if np.mean(scores) > highest_score:
        highest_score = np.mean(scores)
        std = np.std(scores)
        k_features_highest_score = i
    elif np.mean(scores) == highest_score:
        if np.std(scores) < std:
            highest_score = np.mean(scores)
            std = np.std(scores)
            k_features_highest_score = i

# Print the number of features
print('Number of features when highest score: %i' % k_features_highest_score)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\model_selection\_spli
t.py:652: Warning: The least populated class in y has only 1 members, whic
h is too few. The minimum number of members in any class cannot be less th
an n_splits=10.
  % (min_groups, self.n_splits)), Warning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
```

```
c.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\linear_model\logisti
c.py:460: FutureWarning: Default multi_class will be changed to 'auto' in
0.22. Specify the multi_class option to silence this warning.
  "this warning.", FutureWarning)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\svm\base.py:922: Conv
ergenceWarning: Liblinear failed to converge, increase the number of itera
tions.
  "the number of iterations.", ConvergenceWarning)

CV accuracy (original): 0.994 +/- 0.003

---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
<ipython-input-77-5adbcfa86460> in <module>
     10
     11 ## Get score using models with feature selection
---> 12 for i in range(1, X_train_poly.shape[1]+1, 1):
     13     # Select i features
     14     select = SelectKBest(score_func=chi2, k=i)

NameError: name 'X_train_poly' is not defined
```

In [89]:

```
# Select features
select = SelectKBest(score_func=chi2, k=k_features_highest_score)
select.fit(X_train_poly, y_train)
X_train_poly_selected = select.transform(X_train_poly)
```

```
---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
<ipython-input-89-66af41846b7c> in <module>
      1 # Select features
----> 2 select = SelectKBest(score_func=chi2, k=k_features_highest_score)
      3 select.fit(X_train_poly, y_train)
      4 X_train_poly_selected = select.transform(X_train_poly)

NameError: name 'k_features_highest_score' is not defined
```

In [91]:

```
filepath=r"C:\Users\Anustup\Desktop\dataset_malwares.csv"
```

In [92]:

```
malwares = pd.read_csv(filepath, dtype=str)
```

In [93]:

```
print('Found (' + str(len(malwares.index)) + ') malwares in csv file.')
```

```
Found (19611) malwares in csv file.
```

In [103]:

```
malwares.shape
```

Out[103]:

```
(19611, 79)
```

In [105]:

```
malwares.isnull().sum()
```

Out[105]:

```
Name                          0
e_magic                       0
e_cblp                        0
e_cp                          0
e_crlc                        0
e_cparhdr                     0
e_minalloc                    0
e_maxalloc                    0
e_ss                          0
e_sp                          0
e_csum                        0
e_ip                          0
e_cs                          0
e_lfarlc                      0
e_ovno                        0
e_oemid                       0
e_oeminfo                     0
e_lfanew                      0
Machine                       0
NumberOfSections              0
TimeDateStamp                 0
PointerToSymbolTable          0
NumberOfSymbols               0
SizeOfOptionalHeader          0
Characteristics               0
Magic                         0
MajorLinkerVersion            0
MinorLinkerVersion            0
SizeOfCode                    0
SizeOfInitializedData         0
                             ..
SizeOfHeapReserve             0
SizeOfHeapCommit              0
LoaderFlags                   0
NumberOfRvaAndSizes           0
Malware                       0
SuspiciousImportFunctions     0
SuspiciousNameSection         0
SectionsLength                0
SectionMinEntropy             0
SectionMaxEntropy             0
SectionMinRawsize             0
SectionMaxRawsize             0
SectionMinVirtualsize         0
SectionMaxVirtualsize         0
SectionMaxPhysical            0
SectionMinPhysical            0
SectionMaxVirtual             0
SectionMinVirtual             0
SectionMaxPointerData         0
SectionMinPointerData         0
SectionMaxChar                0
SectionMainChar               0
DirectoryEntryImport          0
DirectoryEntryImportSize      0
DirectoryEntryExport          0
ImageDirectoryEntryExport     0
ImageDirectoryEntryImport     0
ImageDirectoryEntryResource   0
```

```
ImageDirectoryEntryException    0
ImageDirectoryEntrySecurity     0
Length: 79, dtype: int64
```

In [107]:

```
malwares.columns
```

Out[107]:

```
Index(['Name', 'e_magic', 'e_cblp', 'e_cp', 'e_crlc', 'e_cparhdr',
       'e_minalloc', 'e_maxalloc', 'e_ss', 'e_sp', 'e_csum', 'e_ip', 'e_c
s',
       'e_lfarlc', 'e_ovno', 'e_oemid', 'e_oeminfo', 'e_lfanew', 'Machin
e',
       'NumberOfSections', 'TimeDateStamp', 'PointerToSymbolTable',
       'NumberOfSymbols', 'SizeOfOptionalHeader', 'Characteristics', 'Magi
c',
       'MajorLinkerVersion', 'MinorLinkerVersion', 'SizeOfCode',
       'SizeOfInitializedData', 'SizeOfUninitializedData',
       'AddressOfEntryPoint', 'BaseOfCode', 'ImageBase', 'SectionAlignmen
t',
       'FileAlignment', 'MajorOperatingSystemVersion',
       'MinorOperatingSystemVersion', 'MajorImageVersion', 'MinorImageVers
ion',
       'MajorSubsystemVersion', 'MinorSubsystemVersion', 'SizeOfHeaders',
       'CheckSum', 'SizeOfImage', 'Subsystem', 'DllCharacteristics',
       'SizeOfStackReserve', 'SizeOfStackCommit', 'SizeOfHeapReserve',
       'SizeOfHeapCommit', 'LoaderFlags', 'NumberOfRvaAndSizes', 'Malwar
e',
       'SuspiciousImportFunctions', 'SuspiciousNameSection', 'SectionsLeng
th',
       'SectionMinEntropy', 'SectionMaxEntropy', 'SectionMinRawsize',
       'SectionMaxRawsize', 'SectionMinVirtualsize', 'SectionMaxVirtualsiz
e',
       'SectionMaxPhysical', 'SectionMinPhysical', 'SectionMaxVirtual',
       'SectionMinVirtual', 'SectionMaxPointerData', 'SectionMinPointerDat
a',
       'SectionMaxChar', 'SectionMainChar', 'DirectoryEntryImport',
       'DirectoryEntryImportSize', 'DirectoryEntryExport',
       'ImageDirectoryEntryExport', 'ImageDirectoryEntryImport',
       'ImageDirectoryEntryResource', 'ImageDirectoryEntryException',
       'ImageDirectoryEntrySecurity'],
      dtype='object')
```

In [108]:

```
data1=malwares.dropna(how="any",axis=0)
data1.head()
```

Out[108]:

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr
0	VirusShare_a878ba26000edaac5c98eff4432723b3	23117	144	3	0	4
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	23117	144	3	0	4
2	VirusShare_ef84cdeba22be72a69b198213dada81a	23117	144	3	0	4
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	23117	144	3	0	4
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	23117	144	3	0	4

5 rows × 79 columns

In [110]:

```
data1["e_magic"].value_counts()
```

Out[110]:

```
23117    19611
Name: e_magic, dtype: int64
```

In [115]:

```
data1.head()
```

Out[115]:

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr
0	VirusShare_a878ba26000edaac5c98eff4432723b3	23117	144	3	0	4
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	23117	144	3	0	4
2	VirusShare_ef84cdeba22be72a69b198213dada81a	23117	144	3	0	4
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	23117	144	3	0	4
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	23117	144	3	0	4

5 rows × 79 columns

In [116]:

```
data1.tail()
```

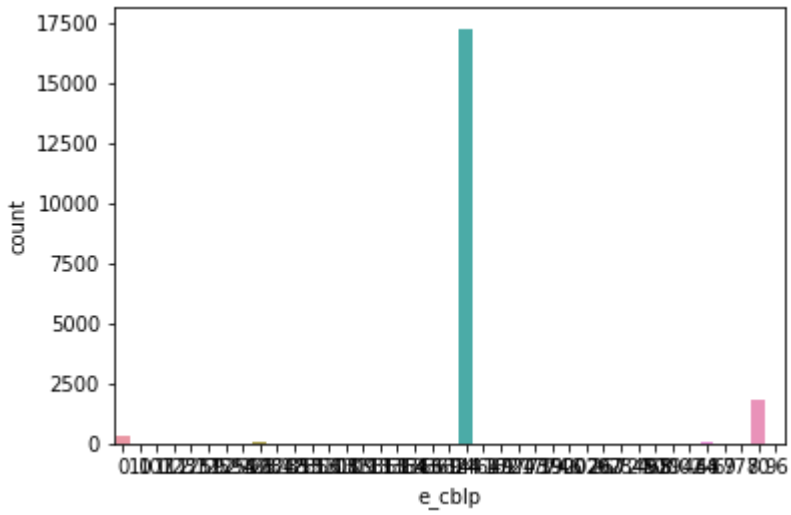
Out[116]:

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr	e_mi
19606	clip.exe	23117	144	3	0	4	
19607	VNC-Server-6.2.0-Windows.exe	23117	144	3	0	4	
19608	Microsoft.GroupPolicy.Management.ni.dll	23117	0	0	0	0	
19609	cryptuiwizard.dll	23117	144	3	0	4	
19610	winhttp.dll	23117	144	3	0	4	

5 rows × 79 columns

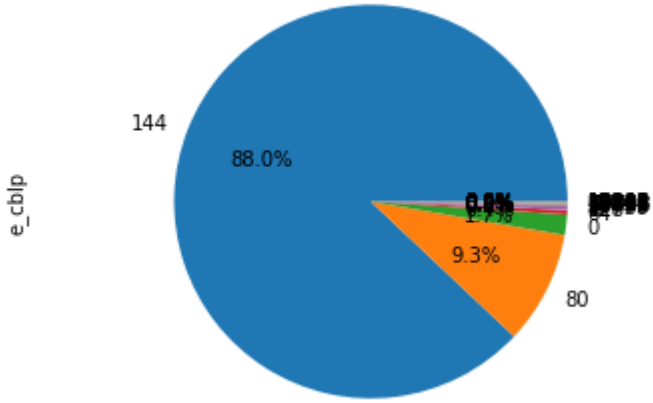
In [119]:

```
sns.countplot(data1["e_cblp"])
plt.show()
```



In [120]:

```
data1["e_cblp"].value_counts().plot(kind="pie",autopct="%1.1f%%")
plt.axis("equal")
plt.show()
```



In [126]:

```
x=data1.drop(["e_cblp","e_magic"],axis=1)
x.head()
```

Out[126]:

	Name	e_cp	e_crlc	e_cparhdr	e_minalloc	e_max
0	VirusShare_a878ba26000edaac5c98eff4432723b3	3	0	4	0	6
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	3	0	4	0	6
2	VirusShare_ef84cdeba22be72a69b198213dada81a	3	0	4	0	6
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	3	0	4	0	6
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	3	0	4	0	6

5 rows × 77 columns

In [128]:

```
y=data1["e_magic"]
y
```

Out[128]:

```
0        23117
1        23117
2        23117
3        23117
4        23117
5        23117
6        23117
7        23117
8        23117
9        23117
10       23117
11       23117
12       23117
13       23117
14       23117
15       23117
16       23117
17       23117
18       23117
19       23117
20       23117
21       23117
22       23117
23       23117
24       23117
25       23117
26       23117
27       23117
28       23117
29       23117
         ...
19581    23117
19582    23117
19583    23117
19584    23117
19585    23117
19586    23117
19587    23117
19588    23117
19589    23117
19590    23117
19591    23117
19592    23117
19593    23117
19594    23117
19595    23117
19596    23117
19597    23117
19598    23117
19599    23117
19600    23117
19601    23117
19602    23117
19603    23117
19604    23117
19605    23117
19606    23117
19607    23117
19608    23117
```

```
19609    23117
19610    23117
Name: e_magic, Length: 19611, dtype: object
```

In [137]:

```
data=pd.read_csv(r"C:\Users\Anustup\Downloads\Malware dataset.csv (3).zip")
```

In [138]:

```
data.head()
```

Out[138]:

	hash	millisecond	classification	state	usage
0	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	0	malware	0	
1	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	1	malware	0	
2	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	2	malware	0	
3	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	3	malware	0	
4	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	4	malware	0	

5 rows × 35 columns

In [140]:

```
data.shape
```

Out[140]:

```
(100000, 35)
```

In [141]:

```
data.isnull().sum()
```

Out[141]:

```
hash                0
millisecond         0
classification      0
state               0
usage_counter       0
prio                0
static_prio         0
normal_prio         0
policy              0
vm_pgoff            0
vm_truncate_count   0
task_size           0
cached_hole_size    0
free_area_cache     0
mm_users            0
map_count           0
hiwater_rss         0
total_vm            0
shared_vm           0
exec_vm             0
reserved_vm         0
nr_ptes             0
end_data            0
last_interval       0
nvcsw               0
nivcsw              0
min_flt             0
maj_flt             0
fs_excl_counter     0
lock                0
utime               0
stime               0
gtime               0
cgtime              0
signal_nvcsw        0
dtype: int64
```

In [142]:

```
data.columns
```

Out[142]:

```
Index(['hash', 'millisecond', 'classification', 'state', 'usage_counter',
       'prio', 'static_prio', 'normal_prio', 'policy', 'vm_pgoff',
       'vm_truncate_count', 'task_size', 'cached_hole_size', 'free_area_ca
che',
       'mm_users', 'map_count', 'hiwater_rss', 'total_vm', 'shared_vm',
       'exec_vm', 'reserved_vm', 'nr_ptes', 'end_data', 'last_interval',
       'nvcsw', 'nivcsw', 'min_flt', 'maj_flt', 'fs_excl_counter', 'lock',
       'utime', 'stime', 'gtime', 'cgtime', 'signal_nvcsw'],
      dtype='object')
```

In [143]:

```
data1=data.dropna(how="any",axis=0)
data1.head()
```

Out[143]:

	hash	millisecond	classification	state	usage
0	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	0	malware	0	
1	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	1	malware	0	
2	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	2	malware	0	
3	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	3	malware	0	
4	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	4	malware	0	

5 rows × 35 columns

In [144]:

```
data1["classification"].value_counts()
```

Out[144]:

```
malware    50000
benign     50000
Name: classification, dtype: int64
```

In [145]:

```
data1['classification'] = data1.classification.map({'benign':0, 'malware':1})
```

In [146]:

```
data.head()
```

Out[146]:

	hash	millisecond	classification	state	usage
0	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	0	malware	0	
1	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	1	malware	0	
2	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	2	malware	0	
3	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	3	malware	0	
4	42fb5e2ec009a05ff5143227297074f1e9c6c3ebb9c914...	4	malware	0	

5 rows × 35 columns

In [147]:

```
data.tail()
```

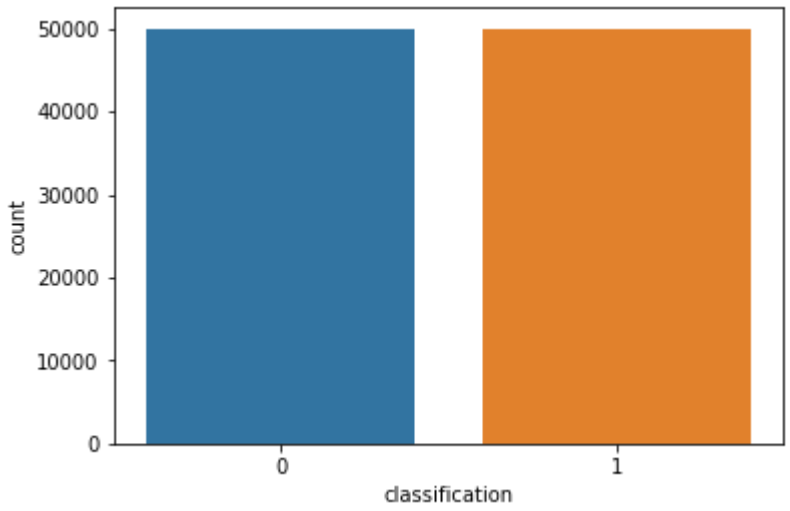
Out[147]:

	hash	millisecond	classification	state
99995	025c63d266e05d9e3bd57dd9ebd0abe904616f569fe4e2...	995	malware	4096
99996	025c63d266e05d9e3bd57dd9ebd0abe904616f569fe4e2...	996	malware	4096
99997	025c63d266e05d9e3bd57dd9ebd0abe904616f569fe4e2...	997	malware	4096
99998	025c63d266e05d9e3bd57dd9ebd0abe904616f569fe4e2...	998	malware	4096
99999	025c63d266e05d9e3bd57dd9ebd0abe904616f569fe4e2...	999	malware	4096

5 rows × 35 columns

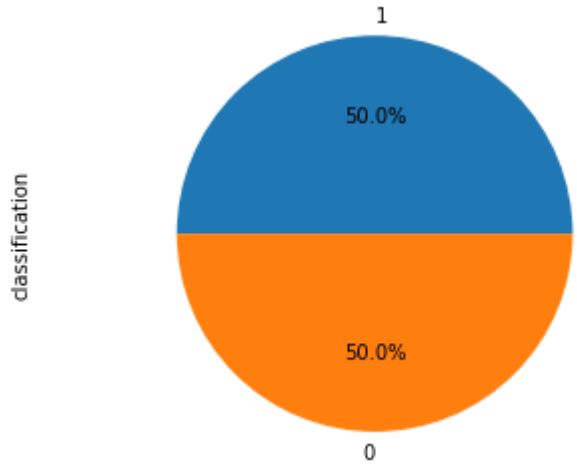
In [148]:

```
sns.countplot(data1["classification"])
plt.show()
```



In [149]:

```
data1["classification"].value_counts().plot(kind="pie",autopct="%1.1f%%")
plt.axis("equal")
plt.show()
```



In [150]:

```
benign1=data.loc[data['classification']=='benign']
benign1["classification"].head()
```

Out[150]:

```
1000    benign
1001    benign
1002    benign
1003    benign
1004    benign
Name: classification, dtype: object
```

In [151]:

```
malware1=data.loc[data['classification']=='malware']
malware1["classification"].head()
```

Out[151]:

```
0    malware
1    malware
2    malware
3    malware
4    malware
Name: classification, dtype: object
```

In [152]:

```
corr=data1.corr()
corr.nlargest(35,'classification')["classification"]
```

Out[152]:

```
classification       1.000000
prio                 0.110036
last_interval        0.006952
min_flt              0.003070
millisecond          0.000000
gtime               -0.014416
stime               -0.042037
free_area_cache     -0.051237
total_vm            -0.059291
state               -0.064702
mm_users            -0.093641
reserved_vm         -0.118608
fs_excl_counter     -0.137883
nivcsw              -0.143791
exec_vm             -0.255123
map_count           -0.271227
static_prio         -0.317941
end_data            -0.324954
maj_flt             -0.324954
shared_vm           -0.324954
vm_truncate_count   -0.354861
utime               -0.369931
nvcsw               -0.386889
Name: classification, dtype: float64
```

In [153]:

```
x=data1.drop(["hash","classification",'vm_truncate_count','shared_vm','exec_vm','nvcsw'
,'maj_flt','utime'],axis=1)
x.head()
```

Out[153]:

	millisecond	state	usage_counter	prio	static_prio	normal_prio	policy	vm_pgoff
0	0	0	0	3069378560	14274	0	0	0
1	1	0	0	3069378560	14274	0	0	0
2	2	0	0	3069378560	14274	0	0	0
3	3	0	0	3069378560	14274	0	0	0
4	4	0	0	3069378560	14274	0	0	0

5 rows × 27 columns

In [154]:

```
y=data1["classification"]
y
```

Out[154]:

```
0        1
1        1
2        1
3        1
4        1
5        1
6        1
7        1
8        1
9        1
10       1
11       1
12       1
13       1
14       1
15       1
16       1
17       1
18       1
19       1
20       1
21       1
22       1
23       1
24       1
25       1
26       1
27       1
28       1
29       1
        ..
99970    1
99971    1
99972    1
99973    1
99974    1
99975    1
99976    1
99977    1
99978    1
99979    1
99980    1
99981    1
99982    1
99983    1
99984    1
99985    1
99986    1
99987    1
99988    1
99989    1
99990    1
99991    1
99992    1
99993    1
99994    1
99995    1
99996    1
99997    1
```

```
99998    1
99999    1
Name: classification, Length: 100000, dtype: int64
```

In [155]:

```
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
```

In [156]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=1)
```

In [157]:

```
from sklearn.naive_bayes import GaussianNB
model=GaussianNB()
model.fit(x_train,y_train)
```

Out[157]:

```
GaussianNB(priors=None, var_smoothing=1e-09)
```

In [158]:

```
pred=model.predict(x_test)
pred
```

Out[158]:

```
array([1, 1, 1, ..., 1, 0, 1], dtype=int64)
```

In [159]:

```
model.score(x_test,y_test)
```

Out[159]:

```
0.6274
```

In [160]:

```
result=pd.DataFrame({
    "Actual_Value":y_test,
    "Predict_Value":pred
})
```

In [161]:

```
result
```

Out[161]:

	Actual_Value	Predict_Value
43660	0	1
87278	1	1
14317	0	1
81932	1	1
95321	1	1
5405	1	1
33188	0	1
63421	1	1
72897	1	1
9507	0	0
88624	1	1
95115	1	1
99243	1	1
77045	1	1
31791	0	1
45417	1	1
71963	1	1
91216	1	1
31924	0	1
15134	0	1
16405	0	1
22718	0	1
15522	0	0
24507	0	1
13979	0	0
71898	1	1
64290	0	0
27706	0	1
92621	1	1
66503	1	1
...	...	...
18845	0	1
64740	0	0
92316	1	1
84568	1	1
9284	0	0
31510	0	1

	Actual_Value	Predict_Value
45911	1	1
7593	0	1
17393	0	1
1407	0	1
30455	0	0
96375	1	1
97553	1	1
54718	0	1
96667	1	1
10506	1	1
37636	0	1
19884	0	0
22766	0	1
13499	0	1
90422	1	1
23841	0	0
24559	0	0
7599	0	1
56585	1	1
994	1	1
42287	0	1
4967	0	1
47725	0	0
42348	0	1

30000 rows × 2 columns

In [12]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import pickle as pck
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
%matplotlib inline
```

In [14]:

```
data = pd.read_csv(r'C:\Users\Anustup\Desktop\Malware Analysis\dataset_malwares.csv', s
ep=',')

#The target is Malware Column {0=Benign, 1=Malware}
X = data.drop(['Name','Malware'], axis=1)
y = data['Malware']

X_train, X_test, y_train, y_test= train_test_split(X,y, test_size=0.2, random_state=101
)
X_train.head()
```

Out[14]:

	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc	e_maxalloc	e_ss	e_sp	e_csı
11441	23117	144	3	0	4	0	65535	0	184	
2624	23117	144	3	0	4	0	65535	0	184	
18874	23117	144	3	0	4	0	65535	0	184	
16415	23117	144	3	0	4	0	65535	0	184	
11179	23117	144	3	0	4	0	65535	0	184	

5 rows × 77 columns

In [15]:

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_train)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\preprocessing\data.p
y:625: DataConversionWarning: Data with input dtype int64, float64 were al
l converted to float64 by StandardScaler.
  return self.partial_fit(X, y)
C:\Users\Anustup\Anaconda3\lib\site-packages\sklearn\base.py:462: DataConv
ersionWarning: Data with input dtype int64, float64 were all converted to
float64 by StandardScaler.
  return self.fit(X, **fit_params).transform(X)
```

In [17]:

```
X_new = pd.DataFrame(X_scaled, columns=X.columns)
X_new.head()
```

Out[17]:

	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc	e_maxalloc	e_ss	
0	0.0	-0.038591	-0.050297	-0.041557	-0.040212	-0.042419	0.148298	-0.016139	-
1	0.0	-0.038591	-0.050297	-0.041557	-0.040212	-0.042419	0.148298	-0.016139	-
2	0.0	-0.038591	-0.050297	-0.041557	-0.040212	-0.042419	0.148298	-0.016139	-
3	0.0	-0.038591	-0.050297	-0.041557	-0.040212	-0.042419	0.148298	-0.016139	-
4	0.0	-0.038591	-0.050297	-0.041557	-0.040212	-0.042419	0.148298	-0.016139	-

5 rows × 77 columns

In [18]:

```
skpca = PCA(n_components=55)
X_pca = skpca.fit_transform(X_new)
print('Variance sum : ', skpca.explained_variance_ratio_.cumsum()[-1])
```

```
Variance sum :  0.9872673777501164
```

In [19]:

```
from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.metrics import classification_report, confusion_matrix
```

In [20]:

```
model = RFC(n_estimators=100, random_state=0,
                          oob_score = True,
                          max_depth = 16,
                          max_features = 'sqrt')
model.fit(X_pca, y_train)

X_test_scaled = scaler.transform(X_test)
X_test_new = pd.DataFrame(X_test_scaled, columns=X.columns)
X_test_pca = skpca.transform(X_test_new)

y_pred = model.predict(X_test_pca)
print(classification_report(y_pred, y_test))
```

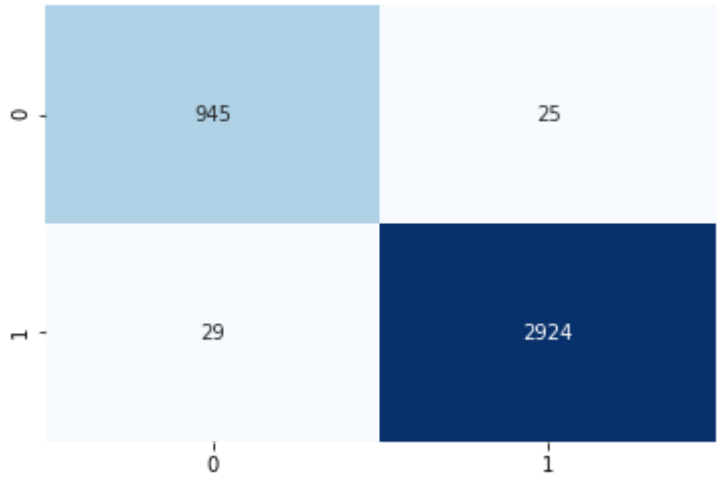
```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97       970
           1       0.99      0.99      0.99      2953

   micro avg       0.99      0.99      0.99      3923
   macro avg       0.98      0.98      0.98      3923
weighted avg       0.99      0.99      0.99      3923


C:\Users\Anustup\Anaconda3\lib\site-packages\ipykernel_launcher.py:7: Data
ConversionWarning: Data with input dtype int64, float64 were all converted
to float64 by StandardScaler.
  import sys
```

In [21]:

```
sns.heatmap(confusion_matrix(y_pred, y_test), annot=True, fmt="d", cmap=plt.cm.Blues, c
bar=False)
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x189f0964da0>
```



In [22]:

```
from sklearn.externals import joblib
from sklearn.pipeline import Pipeline
pipe = Pipeline([('scale', scaler),('pca', skpca), ('clf', model)])
# jbolib.dumps(pipe, 'my_model')
```

In [27]:

```
test = pd.read_csv(r'C:\Users\Anustup\Desktop\Malware Analysis\dataset_malwares.csv', s
ep=',')

X_to_push = test
X_testing = test.drop(['Name'], axis=1)


clf = pipe
X_testing_scaled = clf.named_steps['scale'].transform(X_testing)
X_testing_pca = clf.named_steps['pca'].transform(X_testing_scaled)
y_testing_pred = clf.named_steps['clf'].predict_proba(X_testing_pca)
pd.concat([X_to_push['Name'], pd.DataFrame(y_testing_pred) ], axis=1)
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\ipykernel_launcher.py:8: Data
ConversionWarning: Data with input dtype int64, float64 were all converted
to float64 by StandardScaler.


---------------------------------------------------------------------------
-
ValueError                                Traceback (most recent call las
t)
<ipython-input-27-6909578cf488> in <module>
      6
      7 clf = pipe
----> 8 X_testing_scaled = clf.named_steps['scale'].transform(X_testing)
      9 X_testing_pca = clf.named_steps['pca'].transform(X_testing_scaled)
     10 y_testing_pred = clf.named_steps['clf'].predict_proba(X_testing_pc
a)

~\Anaconda3\lib\site-packages\sklearn\preprocessing\data.py in transform(s
elf, X, y, copy)
    761         else:
    762             if self.with_mean:
--> 763                 X -= self.mean_
    764             if self.with_std:
    765                 X /= self.scale_

ValueError: operands could not be broadcast together with shapes (19611,7
8) (77,) (19611,78)
```

In [28]:

```
from datetime import datetime

print("last update: {}".format(datetime.now()))
```

```
last update: 2019-11-30 12:46:51.392430
```

In [29]:

```
from sklearn.naive_bayes import GaussianNB, BernoulliNB
from sklearn.metrics import accuracy_score, classification_report
from sklearn.ensemble import BaggingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import cohen_kappa_score
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier

from sklearn import preprocessing
import torch
from sklearn import svm
from sklearn import tree
import pandas as pd
from sklearn.externals import joblib
import pickle
import numpy as np
import seaborn as sns
```

```
---------------------------------------------------------------------------
-
ModuleNotFoundError                       Traceback (most recent call las
t)
<ipython-input-29-3a1a65a46fcf> in <module>
     10
     11 from sklearn import preprocessing
---> 12 import torch
     13 from sklearn import svm
     14 from sklearn import tree

ModuleNotFoundError: No module named 'torch'
```

In [30]:

```
import pandas as pd
df = pd.read_csv(r"C:\Users\Anustup\Downloads\datasetandroidpermissions.zip", sep=";")
```

In [31]:

```
df = df.astype("int64")
df.type.value_counts()
```

Out[31]:

```
1    199
0    199
Name: type, dtype: int64
```

In [32]:

```
df.shape
```

Out[32]:

```
(398, 331)
```

In [33]:

```
pd.Series.sort_values(df[df.type==1].sum(axis=0), ascending=False)[1:11]
```

Out[33]:

```
android.permission.INTERNET                 195
android.permission.READ_PHONE_STATE         190
android.permission.ACCESS_NETWORK_STATE     167
android.permission.WRITE_EXTERNAL_STORAGE   136
android.permission.ACCESS_WIFI_STATE        135
android.permission.READ_SMS                 124
android.permission.WRITE_SMS                104
android.permission.RECEIVE_BOOT_COMPLETED   102
android.permission.ACCESS_COARSE_LOCATION     80
android.permission.CHANGE_WIFI_STATE          75
dtype: int64
```

In [34]:

```
pd.Series.sort_values(df[df.type==0].sum(axis=0), ascending=False)[:10]
```

Out[34]:

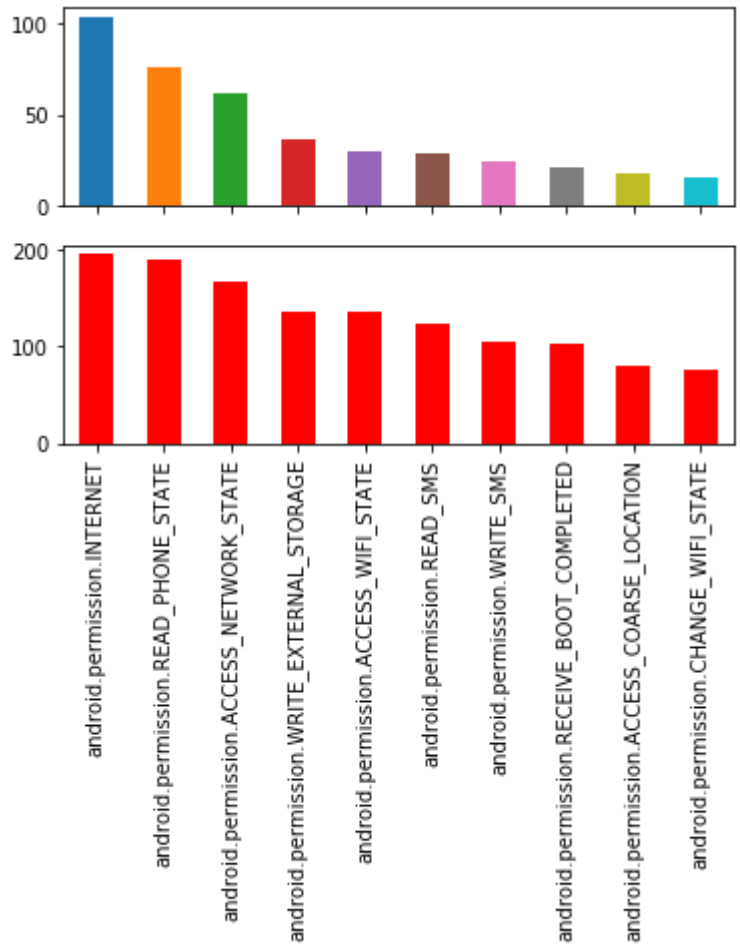
```
android.permission.INTERNET                 104
android.permission.WRITE_EXTERNAL_STORAGE    76
android.permission.ACCESS_NETWORK_STATE      62
android.permission.WAKE_LOCK                 36
android.permission.RECEIVE_BOOT_COMPLETED    30
android.permission.ACCESS_WIFI_STATE         29
android.permission.READ_PHONE_STATE          24
android.permission.VIBRATE                   21
android.permission.ACCESS_FINE_LOCATION      18
android.permission.READ_EXTERNAL_STORAGE     15
dtype: int64
```

In [35]:

```
import matplotlib.pyplot as plt
fig, axs =  plt.subplots(nrows=2, sharex=True)

pd.Series.sort_values(df[df.type==0].sum(axis=0), ascending=False)[:10].plot.bar(ax=axs
[0])
pd.Series.sort_values(df[df.type==1].sum(axis=0), ascending=False)[1:11].plot.bar(ax=ax
s[1], color="red")
```

Out[35]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x189f39da898>
```



In [36]:

```
X_train, X_test, y_train, y_test = train_test_split(df.iloc[:, 1:330], df['type'], test
_size=0.20, random_state=42)
```

In [37]:

```
# Naive Bayes algorithm
gnb = GaussianNB()
gnb.fit(X_train, y_train)

# pred
pred = gnb.predict(X_test)

# accuracy
accuracy = accuracy_score(pred, y_test)
print("naive_bayes")
print(accuracy)
print(classification_report(pred, y_test, labels=None))
```

```
naive_bayes
0.8375
              precision    recall  f1-score   support

           0       0.91      0.76      0.83        41
           1       0.78      0.92      0.85        39

   micro avg       0.84      0.84      0.84        80
   macro avg       0.85      0.84      0.84        80
weighted avg       0.85      0.84      0.84        80
```

In [38]:

```
# kneighbors algorithm

for i in range(3,15,3):

    neigh = KNeighborsClassifier(n_neighbors=i)
    neigh.fit(X_train, y_train)
    pred = neigh.predict(X_test)
    # accuracy
    accuracy = accuracy_score(pred, y_test)
    print("kneighbors {}".format(i))
    print(accuracy)
    print(classification_report(pred, y_test, labels=None))
    print("")
```

```
kneighbors 3
0.8875
              precision    recall  f1-score   support

           0       0.94      0.82      0.88        39
           1       0.85      0.95      0.90        41

   micro avg       0.89      0.89      0.89        80
   macro avg       0.89      0.89      0.89        80
weighted avg       0.89      0.89      0.89        80


kneighbors 6
0.85
              precision    recall  f1-score   support

           0       0.94      0.76      0.84        42
           1       0.78      0.95      0.86        38

   micro avg       0.85      0.85      0.85        80
   macro avg       0.86      0.85      0.85        80
weighted avg       0.87      0.85      0.85        80


kneighbors 9
0.8375
              precision    recall  f1-score   support

           0       0.94      0.74      0.83        43
           1       0.76      0.95      0.84        37

   micro avg       0.84      0.84      0.84        80
   macro avg       0.85      0.85      0.84        80
weighted avg       0.86      0.84      0.84        80


kneighbors 12
0.825
              precision    recall  f1-score   support

           0       0.91      0.74      0.82        42
           1       0.76      0.92      0.83        38

   micro avg       0.82      0.82      0.82        80
   macro avg       0.84      0.83      0.82        80
weighted avg       0.84      0.82      0.82        80
```

In [39]:

```
clf = tree.DecisionTreeClassifier()
clf.fit(X_train, y_train)

# Read the csv test file

pred = clf.predict(X_test)
# accuracy
accuracy = accuracy_score(pred, y_test)
print(clf)
print(accuracy)
print(classification_report(pred, y_test, labels=None))
```

```
---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
<ipython-input-39-56412feb484a> in <module>
----> 1 clf = tree.DecisionTreeClassifier()
      2 clf.fit(X_train, y_train)
      3
      4 # Read the csv test file
      5

NameError: name 'tree' is not defined
```

In [41]:

```
import pandas as pd
data = pd.read_csv(r"C:\Users\Anustup\Downloads\datasetandroidpermissions.zip", sep=";"
)
data.head()
```

Out[41]:

	android	android.app.cts.permission.TEST_GRANTED	android.intent.category.MASTER_CLEAR
0	0	0	
1	0	0	
2	0	0	
3	0	0	
4	0	0	

5 rows × 331 columns

In [45]:

```
data.columns
```

Out[45]:

```
Index(['android', 'android.app.cts.permission.TEST_GRANTED',
       'android.intent.category.MASTER_CLEAR.permission.C2D_MESSAGE',
       'android.os.cts.permission.TEST_GRANTED',
       'android.permission.ACCESS_ALL_DOWNLOADS',
       'android.permission.ACCESS_ALL_EXTERNAL_STORAGE',
       'android.permission.ACCESS_BLUETOOTH_SHARE',
       'android.permission.ACCESS_CACHE_FILESYSTEM',
       'android.permission.ACCESS_CHECKIN_PROPERTIES',
       'android.permission.ACCESS_COARSE_LOCATION',
       ...
       'com.android.voicemail.permission.WRITE_VOICEMAIL',
       'com.foo.mypermission', 'com.foo.mypermission2',
       'org.chromium.chrome.shell.permission.C2D_MESSAGE',
       'org.chromium.chrome.shell.permission.DEBUG',
       'org.chromium.chrome.shell.permission.SANDBOX',
       'org.chromium.chromecast.shell.permission.SANDBOX',
       'org.chromium.content_shell.permission.SANDBOX', 'test_permission',
       'type'],
      dtype='object', length=331)
```

In [46]:

```
data.shape
```

Out[46]:

```
(398, 331)
```

In [47]:

```
data.type.value_counts()
```

Out[47]:

```
1    199
0    199
Name: type, dtype: int64
```

In [48]:

```
data.isna().sum()
```

Out[48]:

```
android                                                    0
android.app.cts.permission.TEST_GRANTED                    0
android.intent.category.MASTER_CLEAR.permission.C2D_MESSAGE    0
android.os.cts.permission.TEST_GRANTED                     0
android.permission.ACCESS_ALL_DOWNLOADS                    0
android.permission.ACCESS_ALL_EXTERNAL_STORAGE             0
android.permission.ACCESS_BLUETOOTH_SHARE                  0
android.permission.ACCESS_CACHE_FILESYSTEM                 0
android.permission.ACCESS_CHECKIN_PROPERTIES               0
android.permission.ACCESS_COARSE_LOCATION                  0
android.permission.ACCESS_CONTENT_PROVIDERS_EXTERNALLY     0
android.permission.ACCESS_DOWNLOAD_MANAGER                 0
android.permission.ACCESS_DOWNLOAD_MANAGER_ADVANCED        0
android.permission.ACCESS_DRM_CERTIFICATES                 0
android.permission.ACCESS_FINE_LOCATION                    0
android.permission.ACCESS_FM_RADIO                         0
android.permission.ACCESS_INPUT_FLINGER                    0
android.permission.ACCESS_KEYGUARD_SECURE_STORAGE          0
android.permission.ACCESS_LOCATION_EXTRA_COMMANDS          0
android.permission.ACCESS_MOCK_LOCATION                    0
android.permission.ACCESS_MTP                              0
android.permission.ACCESS_NETWORK_CONDITIONS               0
android.permission.ACCESS_NETWORK_STATE                    0
android.permission.ACCESS_NOTIFICATIONS                    0
android.permission.ACCESS_PDB_STATE                        0
android.permission.ACCESS_SURFACE_FLINGER                  0
android.permission.ACCESS_WIFI_STATE                       0
android.permission.ACCESS_WIMAX_STATE                      0
android.permission.ACCOUNT_MANAGER                         0
android.permission.ALLOW_ANY_CODEC_FOR_PLAYBACK            0
                                                          ..
com.android.gallery3d.filtershow.permission.WRITE          0
com.android.gallery3d.permission.GALLERY_PROVIDER          0
com.android.launcher.permission.INSTALL_SHORTCUT           0
com.android.launcher.permission.PRELOAD_WORKSPACE          0
com.android.launcher.permission.READ_SETTINGS              0
com.android.launcher.permission.UNINSTALL_SHORTCUT         0
com.android.launcher.permission.WRITE_SETTINGS             0
com.android.launcher3.permission.READ_SETTINGS             0
com.android.launcher3.permission.RECEIVE_FIRST_LOAD_BROADCAST    0
com.android.launcher3.permission.RECEIVE_LAUNCH_BROADCASTS    0
com.android.launcher3.permission.WRITE_SETTINGS            0
com.android.permission.WHITELIST_BLUETOOTH_DEVICE          0
com.android.printspooler.permission.ACCESS_ALL_PRINT_JOBS    0
com.android.providers.tv.permission.ACCESS_ALL_EPG_DATA    0
com.android.providers.tv.permission.ACCESS_WATCHED_PROGRAMS    0
com.android.providers.tv.permission.READ_EPG_DATA          0
com.android.providers.tv.permission.WRITE_EPG_DATA         0
com.android.smspush.WAPPUSH_MANAGER_BIND                   0
com.android.voicemail.permission.ADD_VOICEMAIL             0
com.android.voicemail.permission.READ_VOICEMAIL            0
com.android.voicemail.permission.WRITE_VOICEMAIL           0
com.foo.mypermission                                       0
com.foo.mypermission2                                      0
org.chromium.chrome.shell.permission.C2D_MESSAGE           0
org.chromium.chrome.shell.permission.DEBUG                 0
org.chromium.chrome.shell.permission.SANDBOX               0
org.chromium.chromecast.shell.permission.SANDBOX           0
org.chromium.content_shell.permission.SANDBOX              0
```

```
test_permission                                                 0
type                                                            0
Length: 331, dtype: int64
```

In [49]:

```
data = data.drop(['duracion','avg_local_pkt_rate','avg_remote_pkt_rate'], axis=1).copy
()
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call las
t)
<ipython-input-49-2501129d9930> in <module>
----> 1 data = data.drop(['duracion','avg_local_pkt_rate','avg_remote_pkt_
rate'], axis=1).copy()

~\Anaconda3\lib\site-packages\pandas\core\frame.py in drop(self, labels, a
xis, index, columns, level, inplace, errors)
   3695                                           index=index, columns=co
lumns,
   3696                                           level=level, inplace=in
place,
-> 3697                                           errors=errors)
   3698
   3699     @rewrite_axis_style_signature('mapper', [('copy', True),

~\Anaconda3\lib\site-packages\pandas\core\generic.py in drop(self, labels,
axis, index, columns, level, inplace, errors)
   3109         for axis, labels in axes.items():
   3110             if labels is not None:
-> 3111                 obj = obj._drop_axis(labels, axis, level=level, er
rors=errors)
   3112
   3113         if inplace:

~\Anaconda3\lib\site-packages\pandas\core\generic.py in _drop_axis(self, l
abels, axis, level, errors)
   3141                 new_axis = axis.drop(labels, level=level, errors=e
rrors)
   3142             else:
-> 3143                 new_axis = axis.drop(labels, errors=errors)
   3144             result = self.reindex(**{axis_name: new_axis})
   3145

~\Anaconda3\lib\site-packages\pandas\core\indexes\base.py in drop(self, la
bels, errors)
   4402             if errors != 'ignore':
   4403                 raise KeyError(
-> 4404                     '{} not found in axis'.format(labels[mask]))
   4405             indexer = indexer[~mask]
   4406         return self.delete(indexer)

KeyError: "['duracion' 'avg_local_pkt_rate' 'avg_remote_pkt_rate'] not fou
nd in axis"
```

In [50]:

```
data.describe()
```

Out[50]:

	android	android.app.cts.permission.TEST_GRANTED	android.intent.category.MASTER_CL
count	398.0	398.0	
mean	0.0	0.0	
std	0.0	0.0	
min	0.0	0.0	
25%	0.0	0.0	
50%	0.0	0.0	
75%	0.0	0.0	
max	0.0	0.0	

8 rows × 331 columns

In [74]:

```
import numpy as np, pandas as pd, gc, random
import matplotlib.pyplot as plt
```

In [75]:

```
def load(x):
    ignore = ['MachineIdentifier']
    if x in ignore: return False
    else: return True
```

In [82]:

```
import numpy as np
input_vector = np.array([2, 4, 11])
print(input_vector)
```

```
[ 2  4 11]
```

In [83]:

```
import numpy as np
input_vector = np.array([2, 4, 11])
input_vector = np.array(input_vector, ndmin=2).T
print(input_vector, input_vector.shape)
```

```
[[ 2]
 [ 4]
 [11]] (3, 1)
```

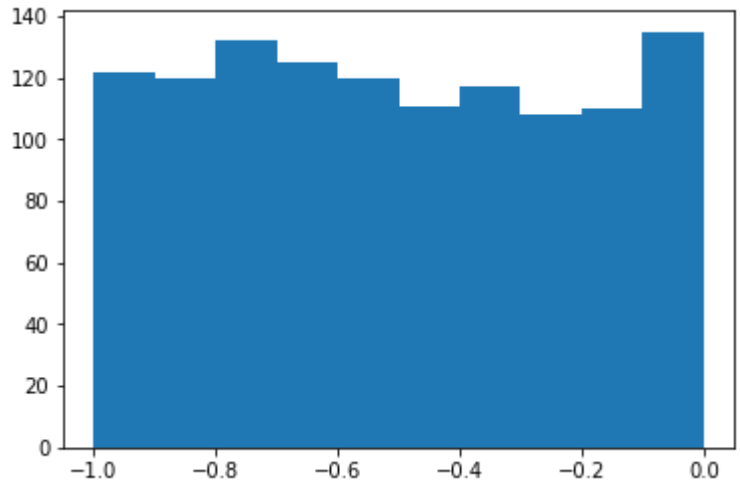
In [84]:

```
import numpy as np
number_of_samples = 1200
low = -1
high = 0
s = np.random.uniform(low, high, number_of_samples)
# all values of s are within the half open interval [-1, 0) :
print(np.all(s >= -1) and np.all(s < 0))
```

```
True
```

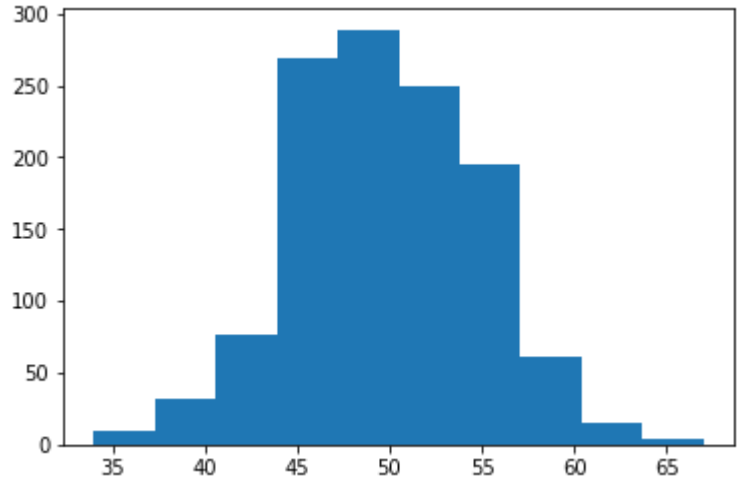
In [85]:

```
import matplotlib.pyplot as plt
plt.hist(s)
plt.show()
```



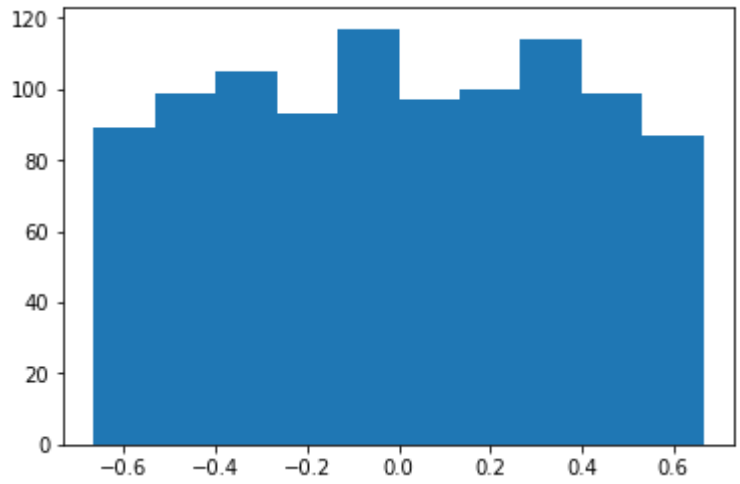
In [86]:

```
s = np.random.binomial(100, 0.5, 1200)
plt.hist(s)
plt.show()
```



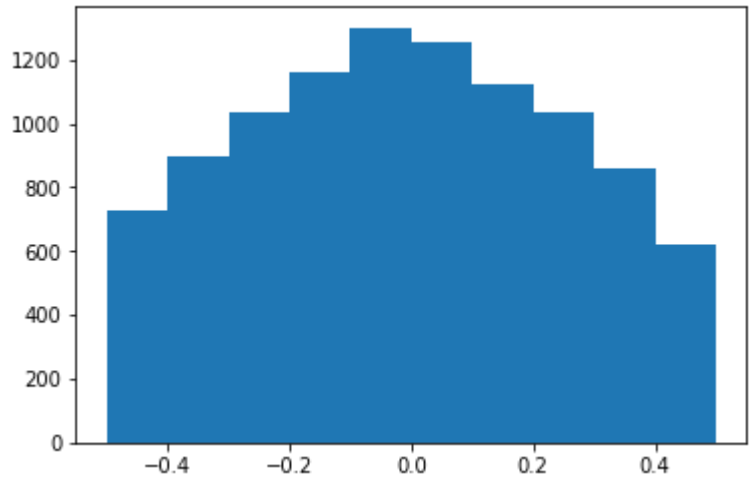
In [87]:

```
from scipy.stats import truncnorm
s = truncnorm(a=-2/3., b=2/3., scale=1, loc=0).rvs(size=1000)
plt.hist(s)
plt.show()
```



In [88]:

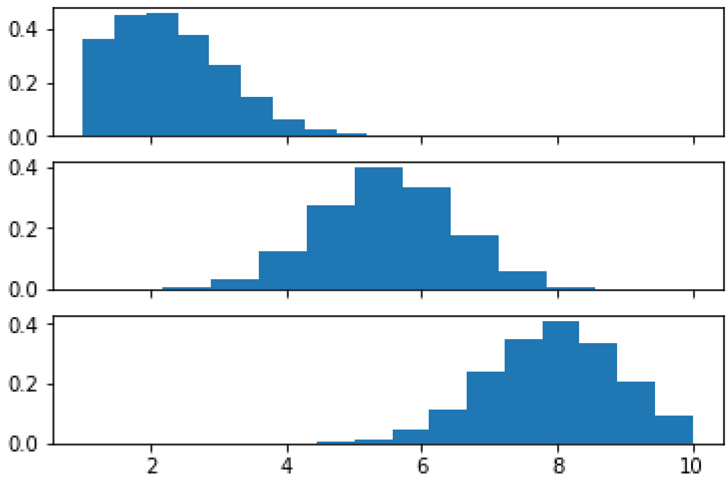
```
def truncated_normal(mean=0, sd=1, low=0, upp=10):
    return truncnorm(
        (low - mean) / sd, (upp - mean) / sd, loc=mean, scale=sd)
X = truncated_normal(mean=0, sd=0.4, low=-0.5, upp=0.5)
s = X.rvs(10000)
plt.hist(s)
plt.show()
```



In [89]:

```
X1 = truncated_normal(mean=2, sd=1, low=1, upp=10)
X2 = truncated_normal(mean=5.5, sd=1, low=1, upp=10)
X3 = truncated_normal(mean=8, sd=1, low=1, upp=10)
import matplotlib.pyplot as plt
fig, ax = plt.subplots(3, sharex=True)
ax[0].hist(X1.rvs(10000), normed=True)
ax[1].hist(X2.rvs(10000), normed=True)
ax[2].hist(X3.rvs(10000), normed=True)
plt.show()
```

```
C:\Users\Anustup\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:652
1: MatplotlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in
3.1. Use 'density' instead.
  alternative="'density'", removal="3.1")
C:\Users\Anustup\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:652
1: MatplotlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in
3.1. Use 'density' instead.
  alternative="'density'", removal="3.1")
C:\Users\Anustup\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:652
1: MatplotlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in
3.1. Use 'density' instead.
  alternative="'density'", removal="3.1")
```



In [90]:

```
no_of_input_nodes = 3
no_of_hidden_nodes = 4
rad = 1 / np.sqrt(no_of_input_nodes)
X = truncated_normal(mean=2, sd=1, low=-rad, upp=rad)
wih = X.rvs((no_of_hidden_nodes, no_of_input_nodes))
wih
```

Out[90]:

```
array([[-0.27560082, -0.06499738, -0.00483244],
       [ 0.27715409,  0.51832982,  0.25987876],
       [ 0.08978611,  0.29920825,  0.15420503],
       [-0.33524818,  0.28018403,  0.46646342]])
```

In [91]:

```
no_of_hidden_nodes = 4
no_of_output_nodes = 2
rad = 1 / np.sqrt(no_of_hidden_nodes)  # this is the input in this layer!
X = truncated_normal(mean=2, sd=1, low=-rad, upp=rad)
who = X.rvs((no_of_output_nodes, no_of_hidden_nodes))
who
```

Out[91]:

```
array([[ 0.14803314,  0.37522044,  0.21490292,  0.12701587],
       [ 0.18803203, -0.42530747,  0.48439636,  0.23722172]])
```

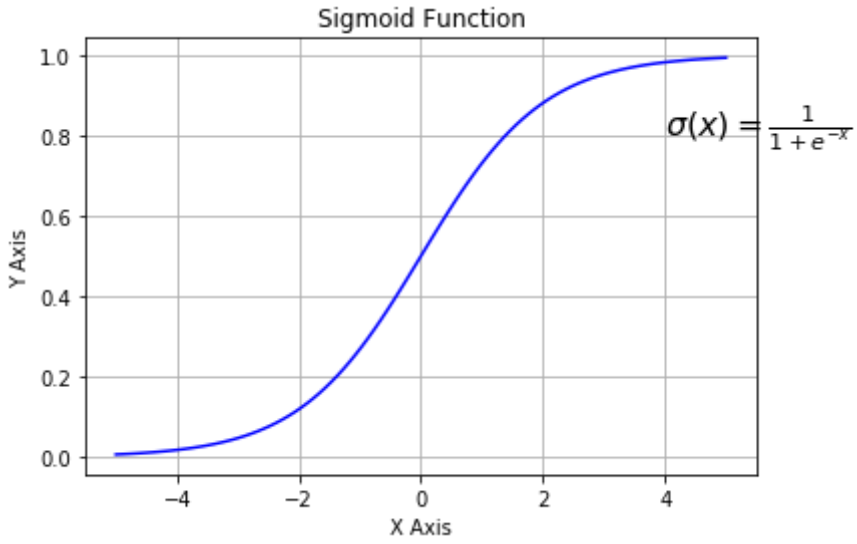
In [92]:

```
class NeuralNetwork:

    def __init__(self,
                 no_of_in_nodes,
                 no_of_out_nodes,
                 no_of_hidden_nodes,
                 learning_rate):
        self.no_of_in_nodes = no_of_in_nodes
        self.no_of_out_nodes = no_of_out_nodes
        self.no_of_hidden_nodes = no_of_hidden_nodes
        self.learning_rate = learning_rate
        self.create_weight_matrices()

    def create_weight_matrices(self):
        rad = 1 / np.sqrt(self.no_of_in_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_in_hidden = X.rvs((self.no_of_hidden_nodes,
                                        self.no_of_in_nodes))
        rad = 1 / np.sqrt(self.no_of_hidden_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_hidden_out = X.rvs((self.no_of_out_nodes,
                                         self.no_of_hidden_nodes))


    def train(self):
        pass

    def run(self):
        pass


if __name__ == "__main__":
    simple_network = NeuralNetwork(no_of_in_nodes = 3,
                                   no_of_out_nodes = 2,
                                   no_of_hidden_nodes = 4,
                                   learning_rate = 0.1)
    print(simple_network.weights_in_hidden)
    print(simple_network.weights_hidden_out)
```

```
[[-0.06971946 -0.21513778  0.46192025]
 [ 0.08867041 -0.34265742 -0.22702858]
 [-0.10199901 -0.55692168 -0.05519091]
 [-0.05000675  0.26244826  0.25785216]]
[[-0.17326096  0.3308556  -0.42680828  0.47948111]
 [-0.00327868  0.02008159  0.05942673  0.11216726]]
```

In [93]:

```
import numpy as np
import matplotlib.pyplot as plt
def sigma(x):
    return 1 / (1 + np.exp(-x))
X = np.linspace(-5, 5, 100)
plt.plot(X, sigma(X),'b')
plt.xlabel('X Axis')
plt.ylabel('Y Axis')
plt.title('Sigmoid Function')
plt.grid()
plt.text(4, 0.8, r'$\sigma(x)=\frac{1}{1+e^{-x}}$', fontsize=16)
plt.show()
```



In [94]:

```
from scipy.special import expit
print(expit(3.4))
print(expit([3, 4, 1]))
print(expit(np.array([0.8, 2.3, 8])))
```

```
0.9677045353015494
[0.95257413 0.98201379 0.73105858]
[0.68997448 0.90887704 0.99966465]
```

In [95]:

```
from scipy.special import expit as activation_function
```

In [96]:

```
from scipy.special import expit as activation_function
from scipy.stats import truncnorm
def truncated_normal(mean=0, sd=1, low=0, upp=10):
    return truncnorm(
        (low - mean) / sd, (upp - mean) / sd, loc=mean, scale=sd)
class NeuralNetwork:

    def __init__(self,
                 no_of_in_nodes,
                 no_of_out_nodes,
                 no_of_hidden_nodes,
                 learning_rate):
        self.no_of_in_nodes = no_of_in_nodes
        self.no_of_out_nodes = no_of_out_nodes
        self.no_of_hidden_nodes = no_of_hidden_nodes
        self.learning_rate = learning_rate
        self.create_weight_matrices()

    def create_weight_matrices(self):
        """ A method to initialize the weight matrices of the neural network"""
        rad = 1 / np.sqrt(self.no_of_in_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_in_hidden = X.rvs((self.no_of_hidden_nodes,
                                        self.no_of_in_nodes))
        rad = 1 / np.sqrt(self.no_of_hidden_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_hidden_out = X.rvs((self.no_of_out_nodes,
                                         self.no_of_hidden_nodes))


    def train(self, input_vector, target_vector):
        pass


    def run(self, input_vector):
        """
        running the network with an input vector input_vector.
        input_vector can be tuple, list or ndarray
        """

        # turning the input vector into a column vector
        input_vector = np.array(input_vector, ndmin=2).T
        output_vector = np.dot(self.weights_in_hidden, input_vector)
        output_vector = activation_function(output_vector)

        output_vector = np.dot(self.weights_hidden_out, output_vector)
        output_vector = activation_function(output_vector)

        return output_vector
```

In [97]:

```
simple_network = NeuralNetwork(no_of_in_nodes=2,
                               no_of_out_nodes=2,
                               no_of_hidden_nodes=10,
                               learning_rate=0.6)
simple_network.run([(3, 4)])
```

Out[97]:

```
array([[0.53487142],
       [0.47011937]])
```

In [98]:

```
@np.vectorize
def sigmoid(x):
    return 1 / (1 + np.e ** -x)
#sigmoid = np.vectorize(sigmoid)
sigmoid([3, 4, 5])
```

Out[98]:

```
array([0.95257413, 0.98201379, 0.99330715])
```

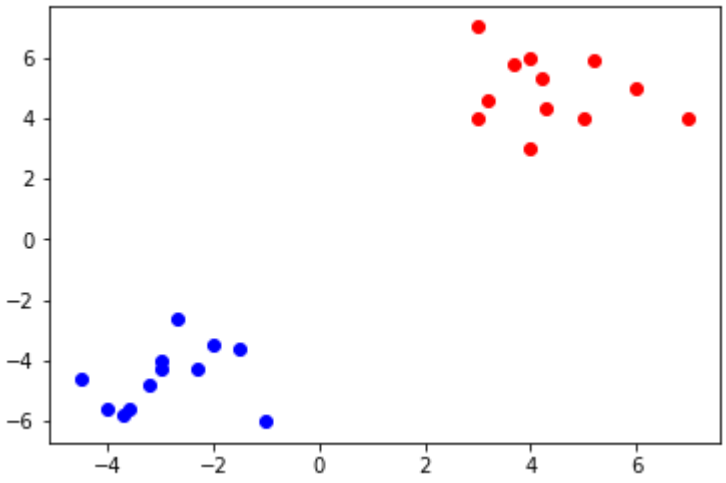
In [99]:

```
import numpy as np
@np.vectorize
def sigmoid(x):
    return 1 / (1 + np.e ** -x)
activation_function = sigmoid
from scipy.stats import truncnorm
def truncated_normal(mean=0, sd=1, low=0, upp=10):
    return truncnorm(
        (low - mean) / sd, (upp - mean) / sd, loc=mean, scale=sd)
class NeuralNetwork:

    def __init__(self,
                 no_of_in_nodes,
                 no_of_out_nodes,
                 no_of_hidden_nodes,
                 learning_rate):
        self.no_of_in_nodes = no_of_in_nodes
        self.no_of_out_nodes = no_of_out_nodes
        self.no_of_hidden_nodes = no_of_hidden_nodes
        self.learning_rate = learning_rate
        self.create_weight_matrices()

    def create_weight_matrices(self):
        """ A method to initialize the weight matrices of the neural network"""
        rad = 1 / np.sqrt(self.no_of_in_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_in_hidden = X.rvs((self.no_of_hidden_nodes,
                                        self.no_of_in_nodes))
        rad = 1 / np.sqrt(self.no_of_hidden_nodes)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_hidden_out = X.rvs((self.no_of_out_nodes,
                                         self.no_of_hidden_nodes))


    def train(self, input_vector, target_vector):
        # input_vector and target_vector can be tuple, list or ndarray

        input_vector = np.array(input_vector, ndmin=2).T
        target_vector = np.array(target_vector, ndmin=2).T

        output_vector1 = np.dot(stelf.weights_in_hidden, input_vector)
        output_vector_hidden = activation_function(output_vector1)

        output_vector2 = np.dot(self.weights_hidden_out, output_vector_hidden)
        output_vector_network = activation_function(output_vector2)

        output_errors = target_vector - output_vector_network
        # update the weights:
        tmp = output_errors * output_vector_network * (1.0 - output_vector_network)
        tmp = self.learning_rate  * np.dot(tmp, output_vector_hidden.T)
        self.weights_hidden_out += tmp
        # calculate hidden errors:
        hidden_errors = np.dot(self.weights_hidden_out.T, output_errors)
        # update the weights:
        tmp = hidden_errors * output_vector_hidden * (1.0 - output_vector_hidden)
        self.weights_in_hidden += self.learning_rate * np.dot(tmp, input_vector.T)


    def run(self, input_vector):
```

```
        # input_vector can be tuple, list or ndarray
        input_vector = np.array(input_vector, ndmin=2).T
        output_vector = np.dot(self.weights_in_hidden, input_vector)
        output_vector = activation_function(output_vector)

        output_vector = np.dot(self.weights_hidden_out, output_vector)
        output_vector = activation_function(output_vector)

        return output_vector
```

In [100]:

```
from matplotlib import pyplot as plt
data1 = [((3, 4), (0.99, 0.01)), ((4.2, 5.3), (0.99, 0.01)),
         ((4, 3), (0.99, 0.01)), ((6, 5), (0.99, 0.01)),
         ((4, 6), (0.99, 0.01)), ((3.7, 5.8), (0.99, 0.01)),
         ((3.2, 4.6), (0.99, 0.01)), ((5.2, 5.9), (0.99, 0.01)),
         ((5, 4), (0.99, 0.01)), ((7, 4), (0.99, 0.01)),
         ((3, 7), (0.99, 0.01)), ((4.3, 4.3), (0.99, 0.01))]
data2 = [((-3, -4), (0.01, 0.99)), ((-2, -3.5), (0.01, 0.99)),
         ((-1, -6), (0.01, 0.99)), ((-3, -4.3), (0.01, 0.99)),
         ((-4, -5.6), (0.01, 0.99)), ((-3.2, -4.8), (0.01, 0.99)),
         ((-2.3, -4.3), (0.01, 0.99)), ((-2.7, -2.6), (0.01, 0.99)),
         ((-1.5, -3.6), (0.01, 0.99)), ((-3.6, -5.6), (0.01, 0.99)),
         ((-4.5, -4.6), (0.01, 0.99)), ((-3.7, -5.8), (0.01, 0.99))]
data = data1 + data2
np.random.shuffle(data)
points1, labels1 = zip(*data1)
X, Y = zip(*points1)
plt.scatter(X, Y, c="r")
points2, labels2 = zip(*data2)
X, Y = zip(*points2)
plt.scatter(X, Y, c="b")
plt.show()
```



In [101]:

```
simple_network = NeuralNetwork(no_of_in_nodes=2,
                               no_of_out_nodes=2,
                               no_of_hidden_nodes=2,
                               learning_rate=0.6)

size_of_learn_sample = int(len(data)*0.9)
learn_data = data[:size_of_learn_sample]
test_data = data[-size_of_learn_sample:]
print()
for i in range(size_of_learn_sample):
    point, label = learn_data[i][0], learn_data[i][1]
    simple_network.train(point, label)

for i in range(size_of_learn_sample):
    point, label = learn_data[i][0], learn_data[i][1]
    cls1, cls2 =simple_network.run(point)
    print(point, cls1, cls2, end=": ")
    if cls1 > cls2:
        if label == (0.99, 0.01):
            print("class1 correct", label)
        else:
            print("class2 incorrect", label)
    else:
        if label == (0.01, 0.99):
            print("class1 correct", label)
        else:
            print("class2 incorrect", label)
```

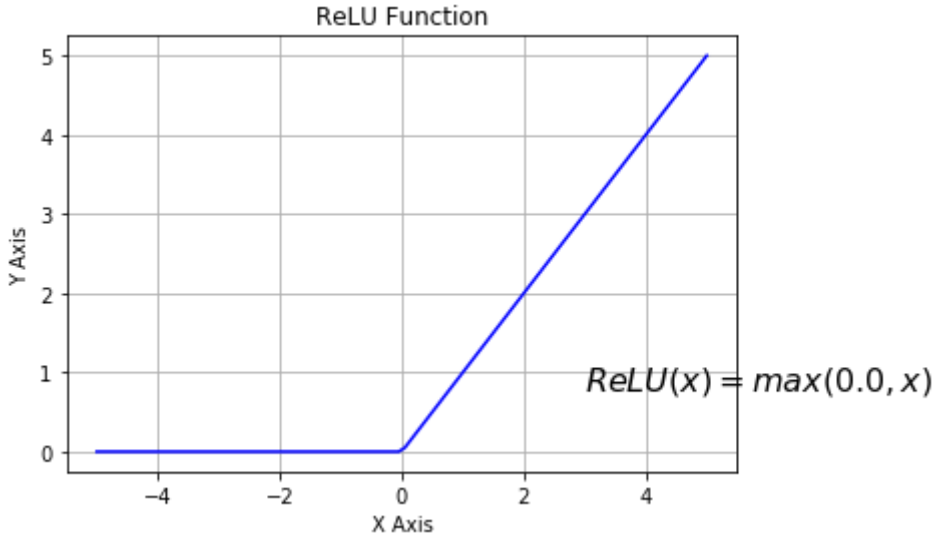
```
---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
<ipython-input-101-db03811285f6> in <module>
     10 for i in range(size_of_learn_sample):
     11     point, label = learn_data[i][0], learn_data[i][1]
---> 12     simple_network.train(point, label)
     13
     14 for i in range(size_of_learn_sample):

<ipython-input-99-494fcf3ce25e> in train(self, input_vector, target_vecto
r)
     39         target_vector = np.array(target_vector, ndmin=2).T
     40
---> 41         output_vector1 = np.dot(stelf.weights_in_hidden, input_vec
tor)
     42         output_vector_hidden = activation_function(output_vector1)
     43

NameError: name 'stelf' is not defined
```

In [102]:

```
# alternative activation function
def ReLU(x):
    return np.maximum(0.0, x)
# derivation of relu
def ReLU_derivation(x):
    if x <= 0:
        return 0
    else:
        return 1
```

In [103]:

```
import numpy as np
import matplotlib.pyplot as plt
X = np.linspace(-5, 5, 100)
plt.plot(X, ReLU(X),'b')
plt.xlabel('X Axis')
plt.ylabel('Y Axis')
plt.title('ReLU Function')
plt.grid()
plt.text(3, 0.8, r'$ReLU(x)=max(0.0, x)$', fontsize=16)
plt.show()
```



In [104]:

```
@np.vectorize
def sigmoid(x):
    return 1 / (1 + np.e ** -x)
activation_function = sigmoid
from scipy.stats import truncnorm
def truncated_normal(mean=0, sd=1, low=0, upp=10):
    return truncnorm(
        (low - mean) / sd, (upp - mean) / sd, loc=mean, scale=sd)
class NeuralNetwork:


    def __init__(self,
                 no_of_in_nodes,
                 no_of_out_nodes,
                 no_of_hidden_nodes,
                 learning_rate,
                 bias=None
                ):
        self.no_of_in_nodes = no_of_in_nodes
        self.no_of_out_nodes = no_of_out_nodes

        self.no_of_hidden_nodes = no_of_hidden_nodes

        self.learning_rate = learning_rate
        self.bias = bias
        self.create_weight_matrices()



    def create_weight_matrices(self):
        """ A method to initialize the weight matrices of the neural
        network with optional bias nodes"""

        bias_node = 1 if self.bias else 0

        rad = 1 / np.sqrt(self.no_of_in_nodes + bias_node)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_in_hidden = X.rvs((self.no_of_hidden_nodes,
                                        self.no_of_in_nodes + bias_node))
        rad = 1 / np.sqrt(self.no_of_hidden_nodes + bias_node)
        X = truncated_normal(mean=0, sd=1, low=-rad, upp=rad)
        self.weights_hidden_out = X.rvs((self.no_of_out_nodes,
                                         self.no_of_hidden_nodes + bias_node))



    def train(self, input_vector, target_vector):
        # input_vector and target_vector can be tuple, list or ndarray

        bias_node = 1 if self.bias else 0
        if self.bias:
            # adding bias node to the end of the inpuy_vector
            input_vector = np.concatenate( (input_vector, [self.bias]) )


        input_vector = np.array(input_vector, ndmin=2).T
        target_vector = np.array(target_vector, ndmin=2).T

        output_vector1 = np.dot(self.weights_in_hidden, input_vector)
```

```
        output_vector_hidden = activation_function(output_vector1)

        if self.bias:
            output_vector_hidden = np.concatenate( (output_vector_hidden, [[self.bias
]]) )


        output_vector2 = np.dot(self.weights_hidden_out, output_vector_hidden)
        output_vector_network = activation_function(output_vector2)

        output_errors = target_vector - output_vector_network
        # update the weights:
        tmp = output_errors * output_vector_network * (1.0 - output_vector_network)
        tmp = self.learning_rate  * np.dot(tmp, output_vector_hidden.T)
        self.weights_hidden_out += tmp
        # calculate hidden errors:
        hidden_errors = np.dot(self.weights_hidden_out.T, output_errors)
        # update the weights:
        tmp = hidden_errors * output_vector_hidden * (1.0 - output_vector_hidden)
        if self.bias:
            x = np.dot(tmp, input_vector.T)[:-1,:]     # ???? last element cut off, ???
        else:
            x = np.dot(tmp, input_vector.T)
        self.weights_in_hidden += self.learning_rate * x



    def run(self, input_vector):
        # input_vector can be tuple, list or ndarray

        if self.bias:
            # adding bias node to the end of the inpuy_vector
            input_vector = np.concatenate( (input_vector, [1]) )
        input_vector = np.array(input_vector, ndmin=2).T
        output_vector = np.dot(self.weights_in_hidden, input_vector)
        output_vector = activation_function(output_vector)

        if self.bias:
            output_vector = np.concatenate( (output_vector, [[1]]) )

        output_vector = np.dot(self.weights_hidden_out, output_vector)
        output_vector = activation_function(output_vector)

        return output_vector
```

In [105]:

```
class1 = [(3, 4), (4.2, 5.3), (4, 3), (6, 5), (4, 6), (3.7, 5.8),
          (3.2, 4.6), (5.2, 5.9), (5, 4), (7, 4), (3, 7), (4.3, 4.3) ]
class2 = [(-3, -4), (-2, -3.5), (-1, -6), (-3, -4.3), (-4, -5.6),
          (-3.2, -4.8), (-2.3, -4.3), (-2.7, -2.6), (-1.5, -3.6),
          (-3.6, -5.6), (-4.5, -4.6), (-3.7, -5.8) ]
labeled_data = []
for el in class1:
    labeled_data.append( [el, [1, 0]])
for el in class2:
    labeled_data.append([el, [0, 1]])

np.random.shuffle(labeled_data)
print(labeled_data[:10])
data, labels = zip(*labeled_data)
labels = np.array(labels)
data = np.array(data)
```

```
[[(4, 3), [1, 0]], [(3.2, 4.6), [1, 0]], [(5, 4), [1, 0]], [(-3.6, -5.6),
[0, 1]], [(-4, -5.6), [0, 1]], [(-3, -4.3), [0, 1]], [(4, 6), [1, 0]], [(-
4.5, -4.6), [0, 1]], [(3, 7), [1, 0]], [(-1.5, -3.6), [0, 1]]]
```

In [106]:

```
simple_network = NeuralNetwork(no_of_in_nodes=2,
                               no_of_out_nodes=2,
                               no_of_hidden_nodes=10,
                               learning_rate=0.1,
                               bias=None)

for _ in range(20):
    for i in range(len(data)):
        simple_network.train(data[i], labels[i])
for i in range(len(data)):
    print(labels[i])
    print(simple_network.run(data[i]))
```

```
[1 0]
[[0.93297381]
 [0.06739553]]
[1 0]
[[0.93627869]
 [0.06397292]]
[1 0]
[[0.93644608]
 [0.06379994]]
[0 1]
[[0.09356425]
 [0.9034806 ]]
[0 1]
[[0.09342102]
 [0.90363702]]
[0 1]
[[0.09589555]
 [0.90100181]]
[1 0]
[[0.9375056 ]
 [0.06269331]]
[0 1]
[[0.09408665]
 [0.90292577]]
[1 0]
[[0.93756289]
 [0.06263875]]
[0 1]
[[0.10443559]
 [0.89223547]]
[0 1]
[[0.10241198]
 [0.89421472]]
[1 0]
[[0.93699808]
 [0.06322787]]
[1 0]
[[0.93745129]
 [0.06274741]]
[1 0]
[[0.93652545]
 [0.06371616]]
[0 1]
[[0.0934056 ]
 [0.90365124]]
[0 1]
[[0.09631627]
 [0.90071624]]
[1 0]
[[0.93502965]
 [0.06526336]]
[1 0]
[[0.93737998]
 [0.06282647]]
[1 0]
[[0.93762857]
 [0.06255992]]
[0 1]
[[0.09716805]
 [0.89968047]]
[0 1]
```

```
[[0.0967921 ]
 [0.90005276]]
[0 1]
[[0.09465309]
 [0.90231852]]
[0 1]
[[0.10832495]
 [0.88791185]]
[1 0]
[[0.93728107]
 [0.06292756]]
```

In [2]:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the fil
es in the input directory

from subprocess import check_output


# Any results you write to the current directory are saved as output.
```

In [3]:

```
df = pd.read_csv(r"C:\Users\Anustup\Desktop\cs448b_ipasn.csv")
df.head(2)
```

Out[3]:

	date	l_ipn	r_asn	f
0	2006-07-01	0	701	1
1	2006-07-01	0	714	1

In [4]:

```
df['date']= pd.to_datetime(df['date'])
df = df.groupby(['date','l_ipn'],as_index=False).sum()
```

In [5]:

```
df['yday'] = df['date'].dt.dayofyear
df['wday'] = df['date'].dt.dayofweek
```

In [6]:

```
ip0 = df[df['l_ipn']==0]
max0 = np.max(ip0['f'])
ip1 = df[df['l_ipn']==1]
max1 = np.max(ip1['f'])
ip2 = df[df['l_ipn']==2]
max2 = np.max(ip2['f'])
ip3 = df[df['l_ipn']==3]
max3 = np.max(ip3['f'])
ip4 = df[df['l_ipn']==4]
max4 = np.max(ip4['f'])
ip5 = df[df['l_ipn']==5]
max5 = np.max(ip5['f'])
ip6 = df[df['l_ipn']==6]
max6 = np.max(ip6['f'])
ip7 = df[df['l_ipn']==7]
max7 = np.max(ip7['f'])
ip8 = df[df['l_ipn']==8]
max8 = np.max(ip8['f'])
ip9 = df[df['l_ipn']==9]
max9 = np.max(ip9['f'])
ip0.head(2)
```

Out[6]:

	date	l_ipn	r_asn	f	yday	wday
0	2006-07-01	0	436704	106	182	5
10	2006-07-02	0	460025	920	183	6

In [7]:

```
count, division = np.histogram(ip0['f'],bins=10)
division
```

Out[7]:

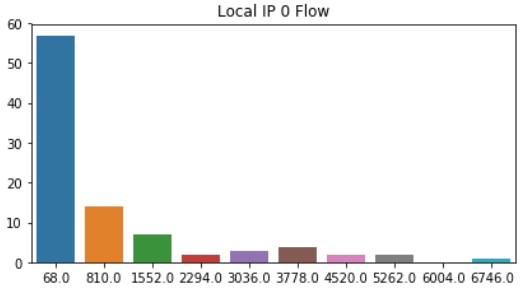
```
array([  68.,  810., 1552., 2294., 3036., 3778., 4520., 5262., 6004.,
       6746., 7488.])
```

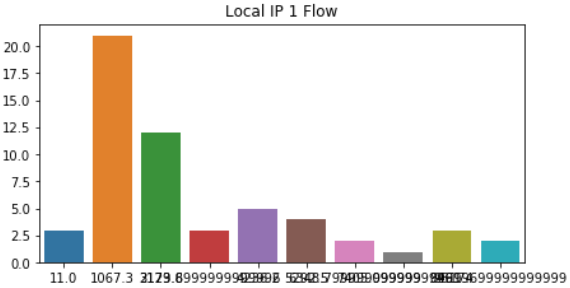
In [8]:

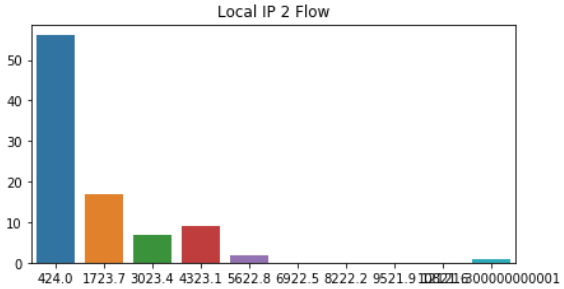
```
f,axarray = plt.subplots(5,2,figsize=(15,20))
count, division = np.histogram(ip0['f'],bins=10)
g = sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[0,0])
axarray[0,0].set_title("Local IP 0 Flow")

count, division = np.histogram(ip1['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[0,1])
axarray[0,1].set_title("Local IP 1 Flow")

count, division = np.histogram(ip2['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[1,0])
axarray[1,0].set_title("Local IP 2 Flow")

count, division = np.histogram(ip3['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[1,1])
axarray[1,1].set_title("Local IP 3 Flow")

count, division = np.histogram(ip4['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[2,0])
axarray[2,1].set_title("Local IP 4 Flow")

count, division = np.histogram(ip5['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[2,1])
axarray[2,1].set_title("Local IP 5 Flow")

count, division = np.histogram(ip6['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[3,0])
axarray[3,0].set_title("Local IP 6 Flow")

count, division = np.histogram(ip7['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[3,1])
axarray[3,1].set_title("Local IP 7 Flow")

count, division = np.histogram(ip8['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[4,0])
axarray[4,0].set_title("Local IP 8 Flow")

count, division = np.histogram(ip9['f'],bins=10)
sns.barplot(x=division[0:len(division)-1],y=count,ax=axarray[4,1])
axarray[4,1].set_title("Local IP 9 Flow")
```

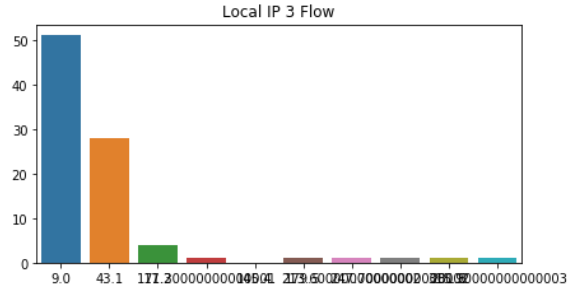
Out[8]:

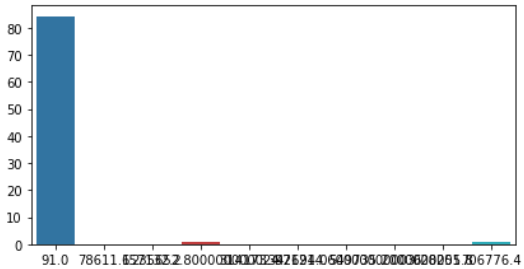
```
Text(0.5, 1.0, 'Local IP 9 Flow')
```

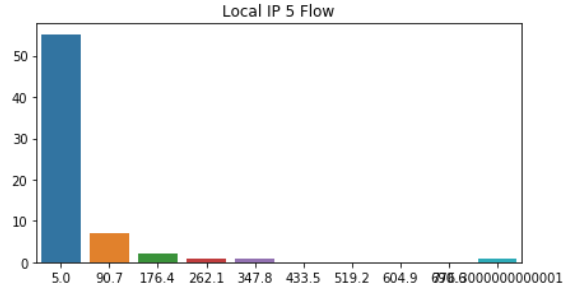

Local IP 0 Flow
68.0 810.0 1552.0 2294.0 3036.0 3778.0 4520.0 5262.0 6004.0 6746.0

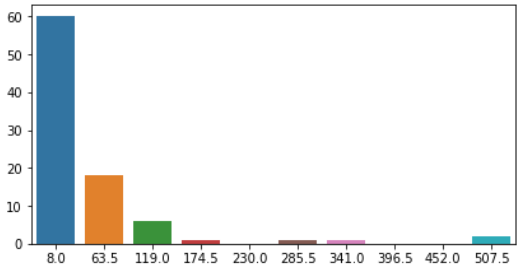

Local IP 1 Flow
11.0

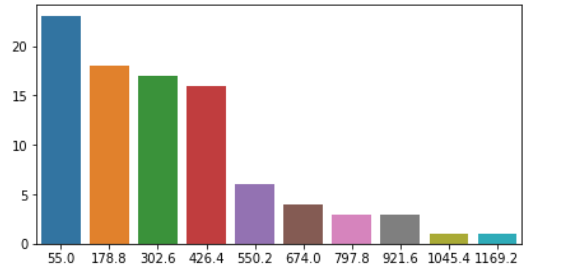

Local IP 2 Flow
424.0 1723.7 3023.4 4323.1 5622.8 6922.5 8222.2 9521.9

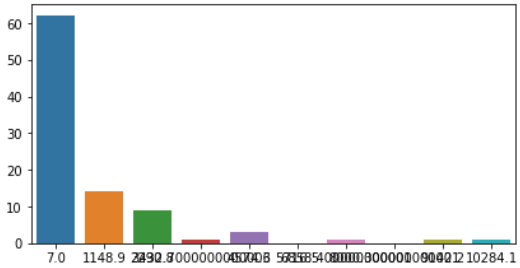

Local IP 3 Flow
9.0 43.1

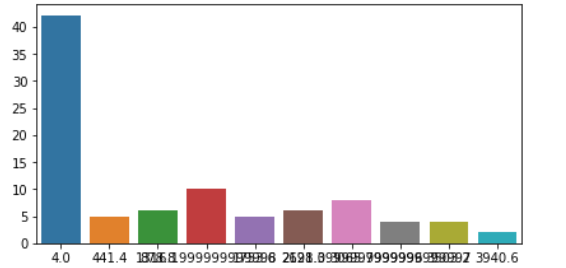

91.0


Local IP 5 Flow
5.0 90.7 176.4 262.1 347.8 433.5 519.2 604.9


Local IP 6 Flow
8.0 63.5 119.0 174.5 230.0 285.5 341.0 396.5 452.0 507.5


Local IP 7 Flow
55.0 178.8 302.6 426.4 550.2 674.0 797.8 921.6 1045.4 1169.2


Local IP 8 Flow
7.0


Local IP 9 Flow
4.0 441.4

In [9]:

```
f,axarray = plt.subplots(5,2,figsize=(15,20))
axarray[0,0].plot(ip0['yday'],ip0['f'])
axarray[0,0].plot(ip0['yday'], [ip0['f'].mean() + 3*ip0['f'].std()]*len(ip0['yday']),co
lor='g')
axarray[0,0].set_title("Local IP 0 Flow")

axarray[0,1].plot(ip1['yday'], ip1['f'])
axarray[0,1].plot(ip1['yday'], [ip1['f'].mean() + 3*ip1['f'].std()]*len(ip1['yday']),co
lor='g')
axarray[0,1].set_title("Local IP 1 Flow")

axarray[1,0].plot(ip2['yday'], ip2['f'])
axarray[1,0].set_title("Local IP 2 Flow")
axarray[1,0].plot(ip2['yday'], [ip2['f'].mean() + 3*ip2['f'].std(ddof=0)]*len(ip2['yda
y']),color='g')


axarray[1,1].plot(ip3['yday'], ip3['f'])
axarray[1,1].set_title("Local IP 3 Flow")
axarray[1,1].plot(ip3['yday'], [ip3['f'].mean() + 3*ip3['f'].std(ddof=0)]*len(ip3['yda
y']),color='g')


axarray[2,0].plot(ip4['yday'], ip4['f'])
axarray[2,0].set_title("Local IP 4 Flow")
axarray[2,0].plot(ip4['yday'], [ip4['f'].mean() + 3*ip4['f'].std(ddof=0)]*len(ip4['yda
y']),color='g')

axarray[2,1].plot(ip5['yday'], ip5['f'])
axarray[2,1].set_title("Local IP 5 Flow")
axarray[2,1].plot(ip5['yday'], [ip5['f'].mean() + 3*ip5['f'].std(ddof=0)]*len(ip5['yda
y']),color='g')

axarray[3,0].plot(ip6['yday'], ip6['f'])
axarray[3,0].set_title("Local IP 6 Flow")
axarray[3,0].plot(ip6['yday'], [ip6['f'].mean() + 3*ip6['f'].std(ddof=0)]*len(ip6['yda
y']),color='g')

axarray[3,1].plot(ip7['yday'], ip7['f'])
axarray[3,1].set_title("Local IP 7 Flow")
axarray[3,1].plot(ip7['yday'], [ip7['f'].mean() + 3*ip7['f'].std(ddof=0)]*len(ip7['yda
y']),color='g')

axarray[4,0].plot(ip8['yday'], ip8['f'])
axarray[4,0].set_title("Local IP 8 Flow")
axarray[4,0].plot(ip8['yday'], [ip8['f'].mean() + 3*ip8['f'].std(ddof=0)]*len(ip8['yda
y']),color='g')


axarray[4,1].plot(ip9['yday'], ip9['f'])
axarray[4,1].set_title("Local IP 9 Flow")
axarray[4,1].plot(ip9['yday'], [ip9['f'].mean() + 3*ip9['f'].std(ddof=0)]*len(ip9['yda
y']),color='g')
```

Out[9]:

```
[<matplotlib.lines.Line2D at 0x2538ffc34e0>]
```

In [10]:

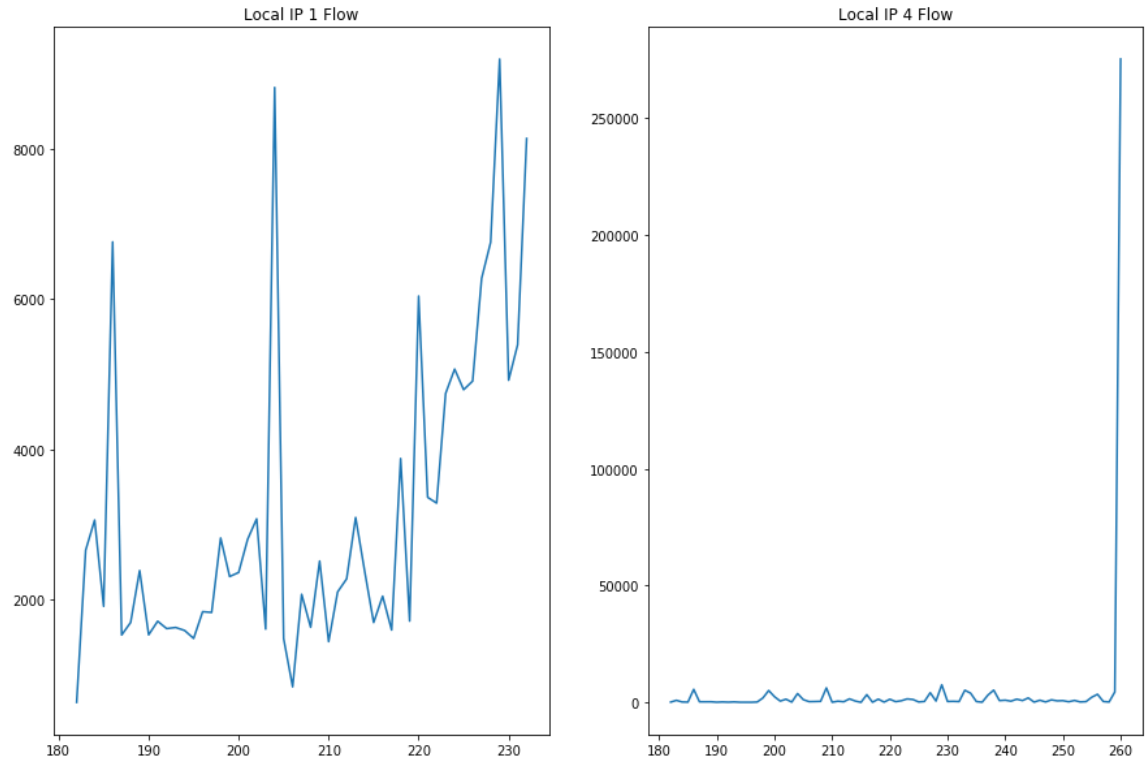
```
ip0 = df[df['l_ipn']==0]
max0 = np.max(ip0['f'])
ip1 = df[df['l_ipn']==1][0:len(ip1['f'])-5]
max1 = np.max(ip1['f'])
ip2 = df[df['l_ipn']==2]
max2 = np.max(ip2['f'])
ip3 = df[df['l_ipn']==3]
max3 = np.max(ip3['f'])
ip4 = df[df['l_ipn']==4][0:len(ip4['f'])-7]
```

In [11]:

```
f,axarray = plt.subplots(1,2,figsize=(15,10))
axarray[0].plot(ip1['yday'],ip1['f'])
axarray[0].set_title("Local IP 1 Flow")
axarray[1].plot(ip4['yday'], ip4['f'])
axarray[1].set_title("Local IP 4 Flow")
```

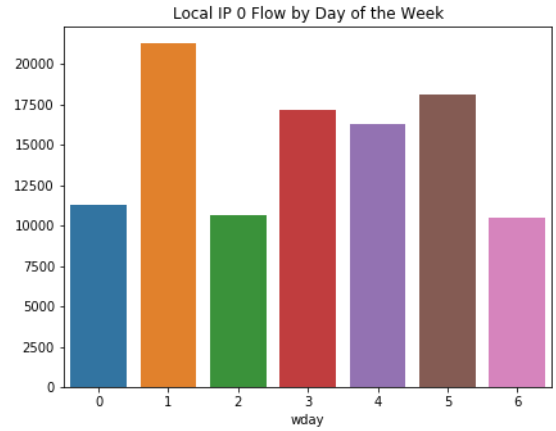
Out[11]:

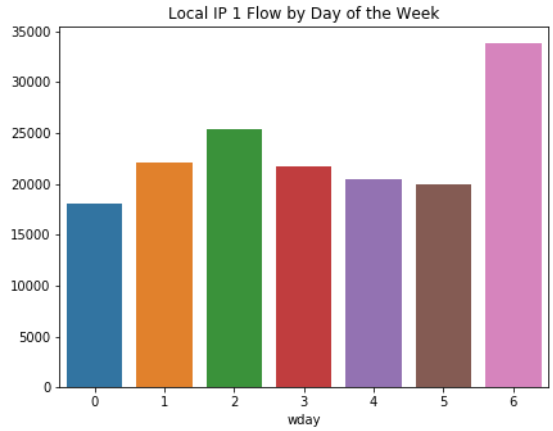
```
Text(0.5, 1.0, 'Local IP 4 Flow')
```

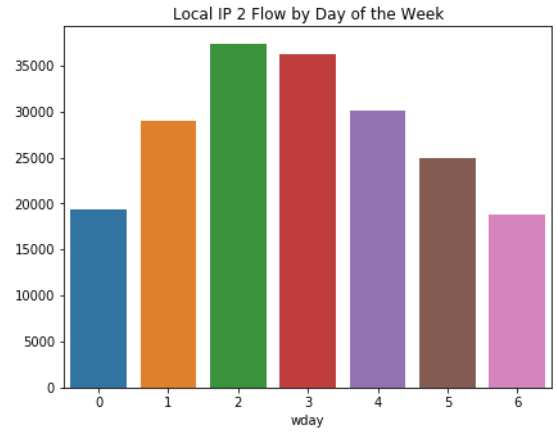


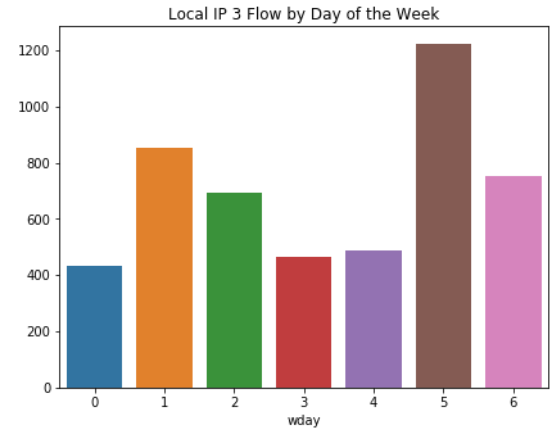
In [12]:

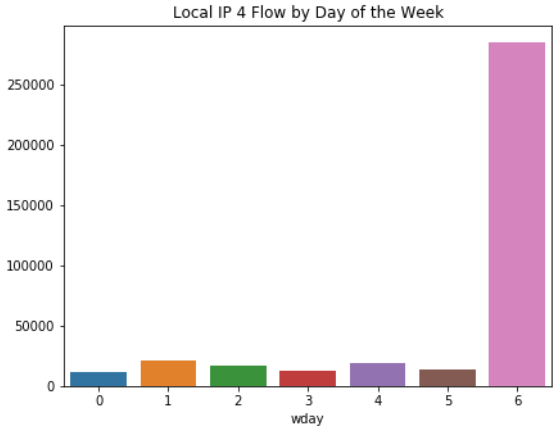
```
f,axarray = plt.subplots(5,2,figsize=(15,30))
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip0.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[0,0])
axarray[0,0].set_title("Local IP 0 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip1.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[0,1])
axarray[0,1].set_title("Local IP 1 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip2.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[1,0])
axarray[1,0].set_title("Local IP 2 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip3.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[1,1])
axarray[1,1].set_title("Local IP 3 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip4.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[2,0])
axarray[2,0].set_title("Local IP 4 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip5.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[2,1])
axarray[2,1].set_title("Local IP 5 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip6.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[3,0])
axarray[3,0].set_title("Local IP 6 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip7.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[3,1])
axarray[3,1].set_title("Local IP 7 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip8.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[4,0])
axarray[4,0].set_title("Local IP 8 Flow by Day of the Week")
sns.barplot(x= ip0.groupby('wday',as_index=False).sum()['wday'],y= ip9.groupby('wday',a
s_index=False).sum()['f'].values,ax=axarray[4,1])
axarray[4,1].set_title("Local IP 9 Flow by Day of the Week")

plt.show()
```

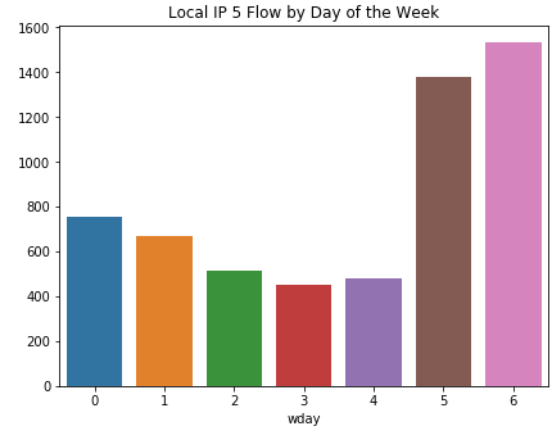

Local IP 0 Flow by Day of the Week
wday

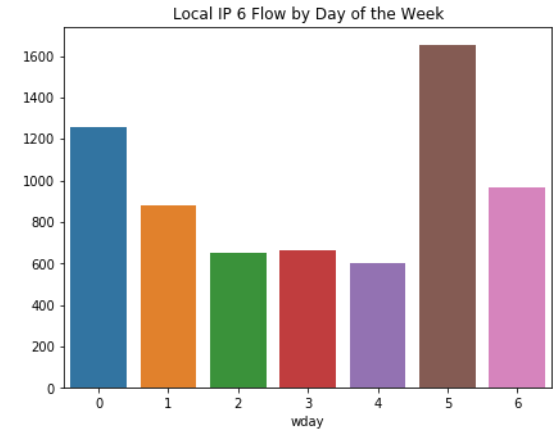

Local IP 1 Flow by Day of the Week
wday

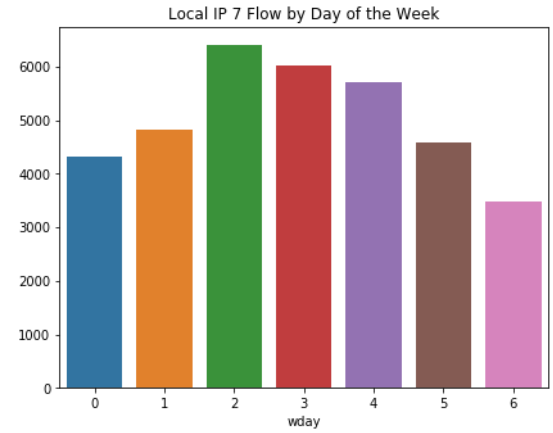

Local IP 2 Flow by Day of the Week
wday

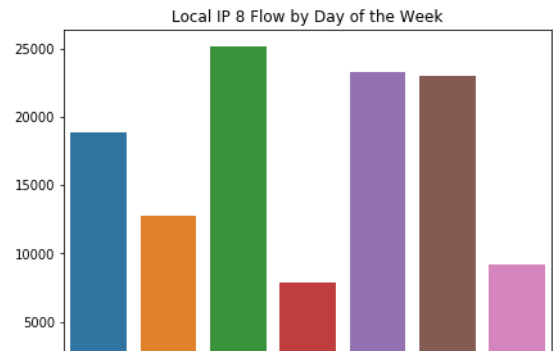

Local IP 3 Flow by Day of the Week
wday

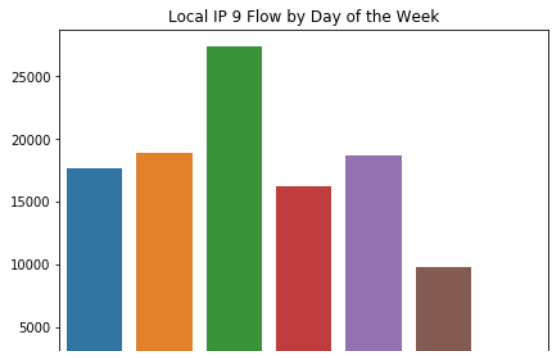

Local IP 4 Flow by Day of the Week
wday


Local IP 5 Flow by Day of the Week
wday


Local IP 6 Flow by Day of the Week
wday


Local IP 7 Flow by Day of the Week
wday


Local IP 8 Flow by Day of the Week


Local IP 9 Flow by Day of the Week

In [13]:

```
plt.plot(range(len(ip0['f'])),ip0['f'].rolling(3).mean()
```

```
  File "<ipython-input-13-86adcccf0608>", line 1
    plt.plot(range(len(ip0['f'])),ip0['f'].rolling(3).mean()
                                                            ^
SyntaxError: unexpected EOF while parsing
```

In [15]:

```
ip0 = df[df['l_ipn']==0]
ip1 = df[df['l_ipn']==1][0:len(df[df['l_ipn']==1])-5]
ip2 = df[df['l_ipn']==2]
ip3 = df[df['l_ipn']==3]
ip4 = df[df['l_ipn']==4][0:len(df[df['l_ipn']==4])-7]

ip5 = df[df['l_ipn']==5]
ip6 = df[df['l_ipn']==6]
ip7 = df[df['l_ipn']==7]
ip8 = df[df['l_ipn']==8]
ip9 = df[df['l_ipn']==9]
```

In [17]:

```
def ApEn(U, m, r):

    def _maxdist(x_i, x_j):
        return max([abs(ua - va) for ua, va in zip(x_i, x_j)])

    def _phi(m):
        x = [[U[j] for j in range(i, i + m - 1 + 1)] for i in range(N - m + 1)]
        C = [len([1 for x_j in x if _maxdist(x_i, x_j) <= r]) / (N - m + 1.0) for x_i i
n x]
        return (N - m + 1.0)**(-1) * sum(np.log(C))

    N = len(U)

    return abs(_phi(m + 1) - _phi(m))
```

In [18]:

```
m=2
r = 3
e0 = ApEn(np.multiply(ip0['f'].values,1),m,r)
e1 = ApEn(np.multiply(ip1['f'].values,1),m,r)
e2 = ApEn(np.multiply(ip2['f'].values,1),m,r)
e3 = ApEn(np.multiply(ip3['f'].values,1),m,r)
e4 = ApEn(np.multiply(ip4['f'].values,1),m,r)
e5 = ApEn(np.multiply(ip5['f'].values,1),m,r)
e6 = ApEn(np.multiply(ip6['f'].values,1),m,r)
e7 = ApEn(np.multiply(ip7['f'].values,1),m,r)
e8 = ApEn(np.multiply(ip8['f'].values,1),m,r)
e9 = ApEn(np.multiply(ip9['f'].values,1),m,r)
```

In [19]:

```
ent_values = pd.DataFrame({'e0':[e0], 'e1':[e1],'e2':[e2],'e3':[e3],'e4':[e4],'e5':[e5
],
              'e6':[e6],'e7':[e7],'e8':[e8],'e9':[e9]})
ent_values.head()
```

Out[19]:

	e0	e1	e2	e3	e4	e5	e6	e7	e8	
0	0.01105	0.020203	0.01105	0.360497	0.012903	0.169414	0.286478	0.004184	0.01105	0.2

In [20]:

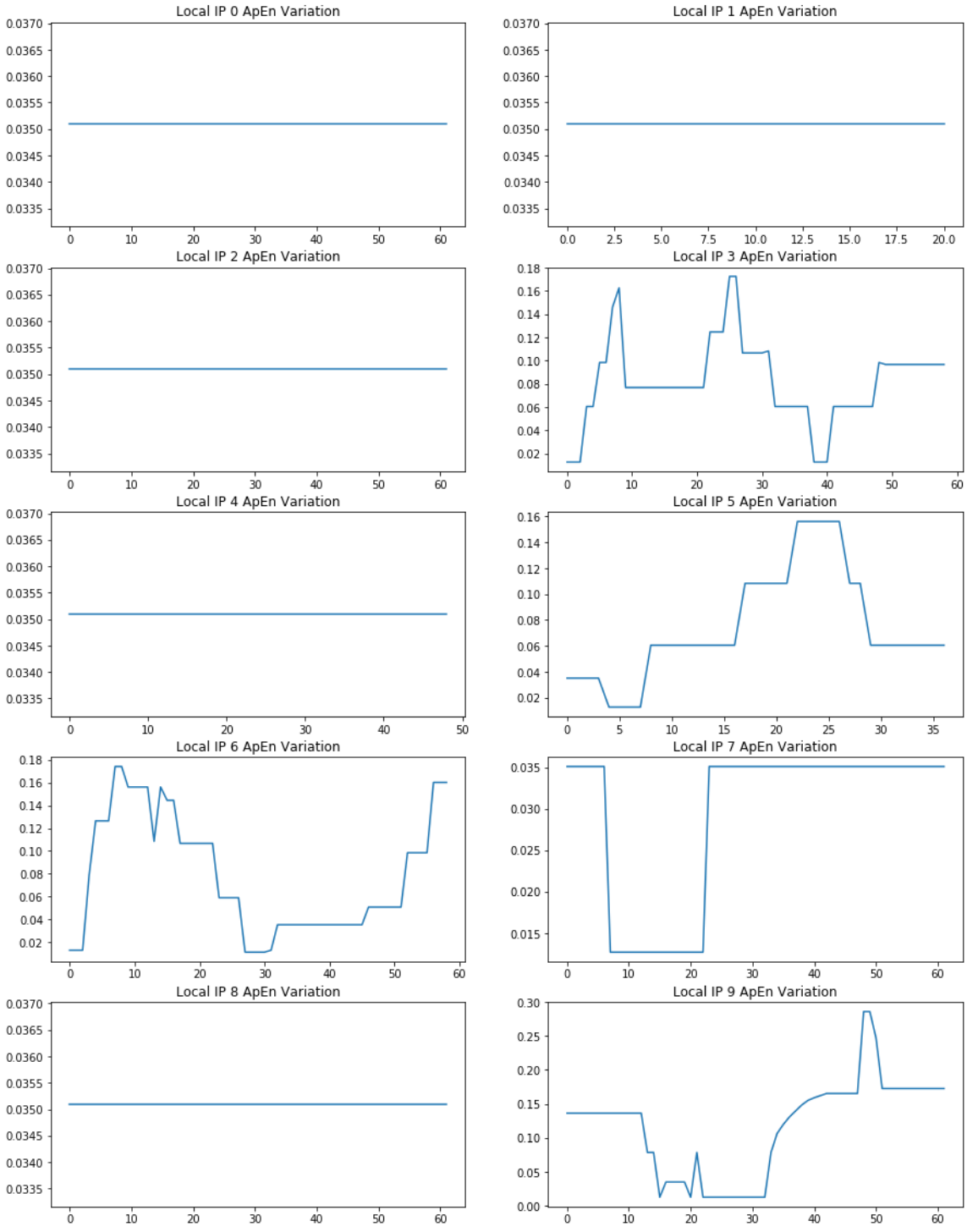
```
def entropyTrend(data,d):
    etrend = [ApEn(np.multiply(data[n:n+d].values,1),2,3) for n in range(len(data)-d)]
    return etrend
```

In [21]:

```
f,axarray = plt.subplots(5,2,figsize=(15,20))
days = 30
et0 = entropyTrend(ip0['f'],days)
axarray[0,0].plot(range(len(et0)),et0)
axarray[0,0].set_title("Local IP 0 ApEn Variation")

et1 = entropyTrend(ip1['f'],days)
axarray[0,1].plot(range(len(et1)),et1)
axarray[0,1].set_title("Local IP 1 ApEn Variation")

et2 = entropyTrend(ip2['f'],days)
axarray[1,0].plot(range(len(et2)),et2)
axarray[1,0].set_title("Local IP 2 ApEn Variation")

et3 = entropyTrend(ip3['f'],days)
axarray[1,1].plot(range(len(et3)),et3)
axarray[1,1].set_title("Local IP 3 ApEn Variation")

et4 = entropyTrend(ip4['f'],days)
axarray[2,0].plot(range(len(et4)),et4)
axarray[2,0].set_title("Local IP 4 ApEn Variation")

et5 = entropyTrend(ip5['f'],days)
axarray[2,1].plot(range(len(et5)),et5)
axarray[2,1].set_title("Local IP 5 ApEn Variation")

et6 = entropyTrend(ip6['f'],days)
axarray[3,0].plot(range(len(et6)),et6)
axarray[3,0].set_title("Local IP 6 ApEn Variation")

et7 = entropyTrend(ip7['f'],days)
axarray[3,1].plot(range(len(et7)),et7)
axarray[3,1].set_title("Local IP 7 ApEn Variation")

et8 = entropyTrend(ip8['f'],days)
axarray[4,0].plot(range(len(et8)),et8)
axarray[4,0].set_title("Local IP 8 ApEn Variation")

et9 = entropyTrend(ip9['f'],days)
axarray[4,1].plot(range(len(et9)),et9)
axarray[4,1].set_title("Local IP 9 ApEn Variation")
```

Out[21]:

```
Text(0.5, 1.0, 'Local IP 9 ApEn Variation')
```



In [ ]:

In [ ]:

In [ ]: