EDA IMAGE

November 15, 2020

1 EDA on Original DATA

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
```

EDA IS DONE WITH TEST DATA ONLY CUZ THAT IS ONLY ATTACKED

[2]:
```
x_test = np.load("./DATA/X_test.npy", allow_pickle=True)
y_test = np.load("./DATA/y_test.npy", allow_pickle=True)
```

[3]:
```
y_test.shape
```

[3]:
```
(5000, 1)
```

[4]:
```
features = np.load("./DATA/FeatureList.npy", allow_pickle=True)
coeff = np.load("./DATA/coeff_features.npy", allow_pickle=True)
```

[5]:
```
x_test_1d = x_test.reshape(x_test.shape[0],-1)
features_1d = features.reshape(-1)
coeff_1d = coeff.reshape(-1)
```

[6]:
```
print(x_test_1d.shape,features_1d.shape,coeff_1d.shape)
```

```
(5000, 2304) (2304,) (2304,)
```

[7]:
```
features.shape
```

[7]:
```
(48, 48)
```

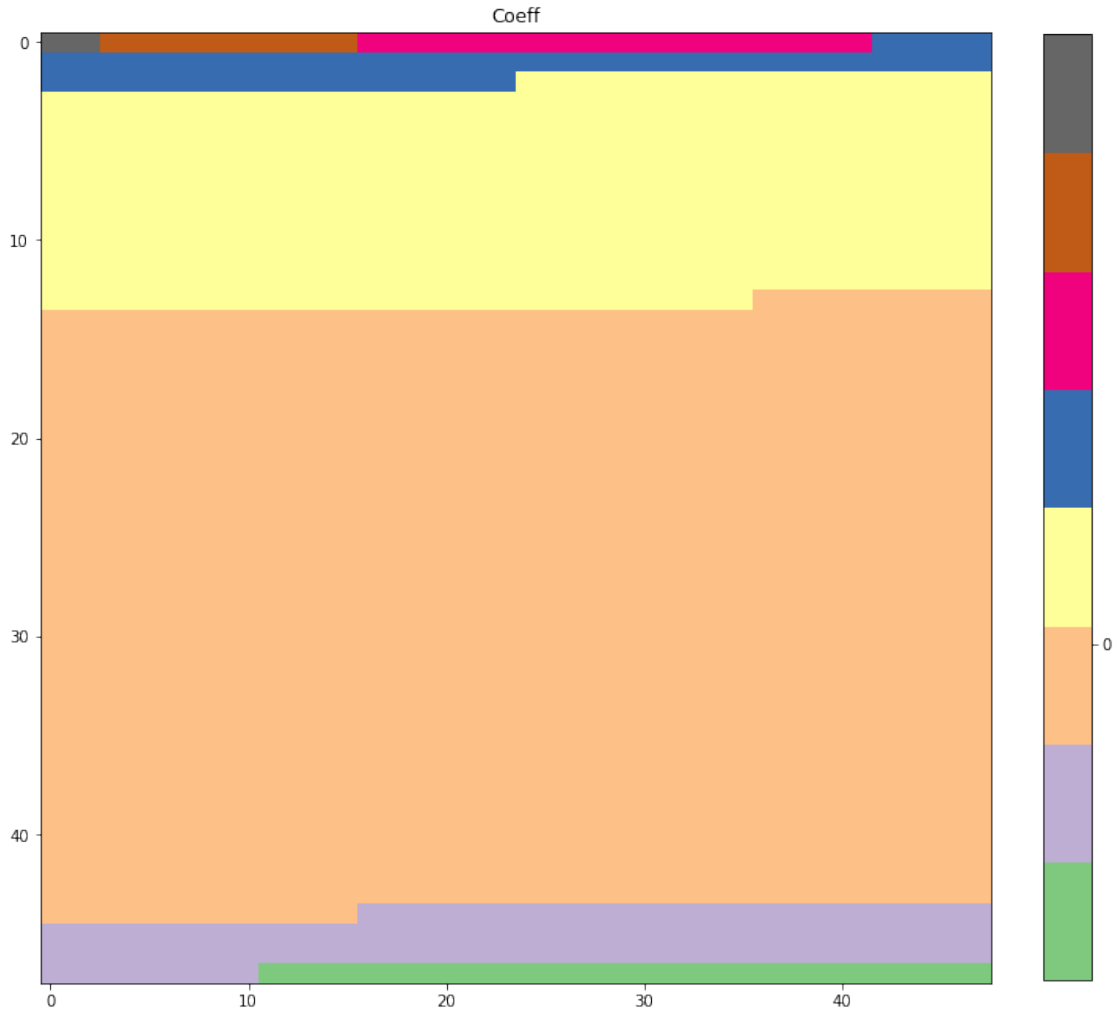
[8]:
```
coeff.shape
```

[8]:
```
(48, 48)
```

[9]:
```
min(coeff.reshape(-1,1))
```

[9]:
```
array([-0.48506692])
```

[24]:
```
fig, ax = plt.subplots(figsize = (12,12))
ax.set_title('Coeff')
cax = ax.imshow(coeff, cmap = plt.cm.Accent)

cbar = plt.colorbar(cax, ticks=[0, 1, 2, 3, 4, 5, 6, 7],
                    orientation='vertical',
                    fraction=0.045, pad=0.05)
```



The above density 2d array of coeffs plotted shows us that most MAL features are grayish(top left) and most BEN features at the bottom right

1.0.1 Find MAL and BEN Fets

[11]:
```
MAL_F = []
BEN_F = []
for i in range(48*48):
    if coeff.reshape(-1,1)[i] > 0:
        MAL_F.append(i)
    else:
        BEN_F.append(i)
print(len(MAL_F),len(BEN_F))
```

```
1329 975
```

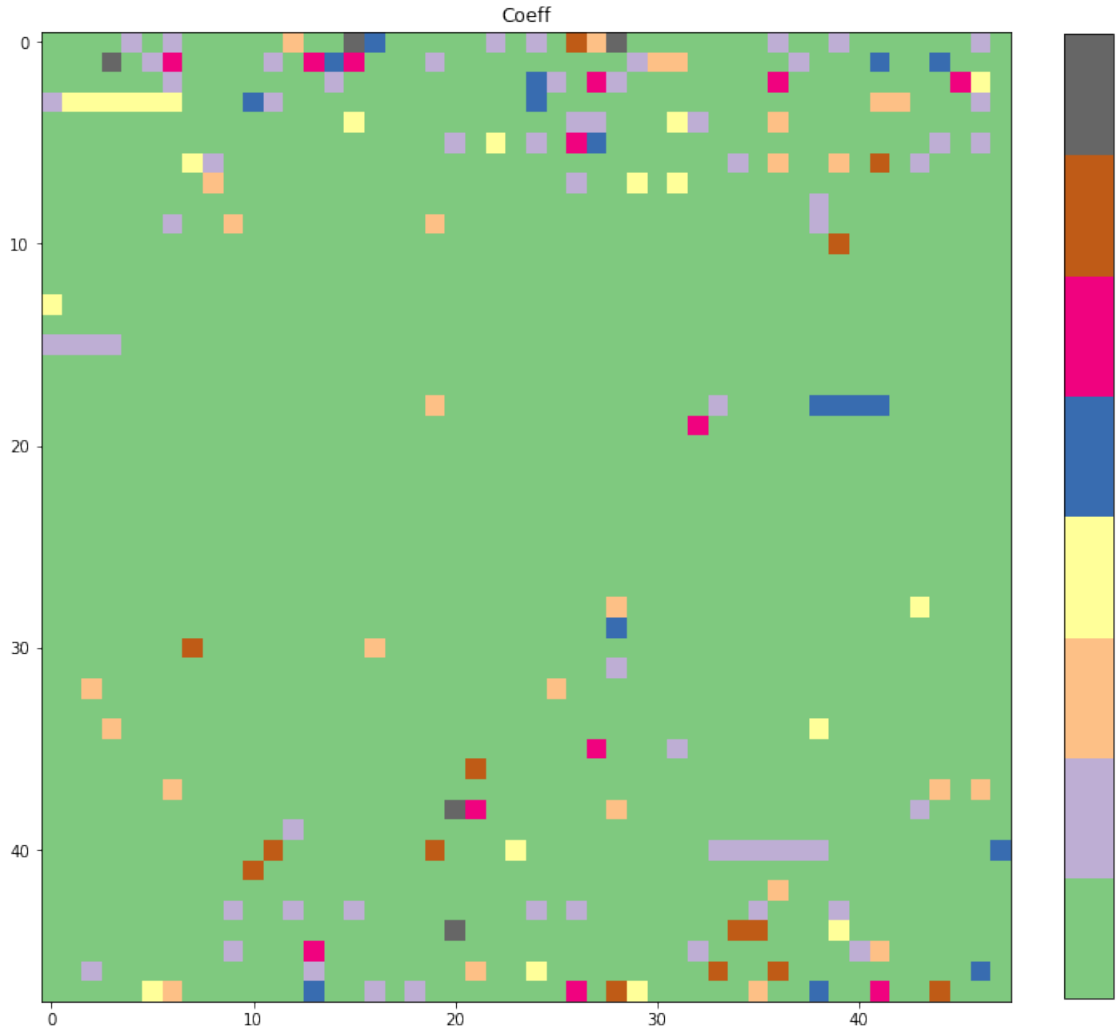
We have a lot of MAL features compared to BEN features but we did choose the features statistically so not touching that

2 Overall Dataset

2.1 Frequency of Features

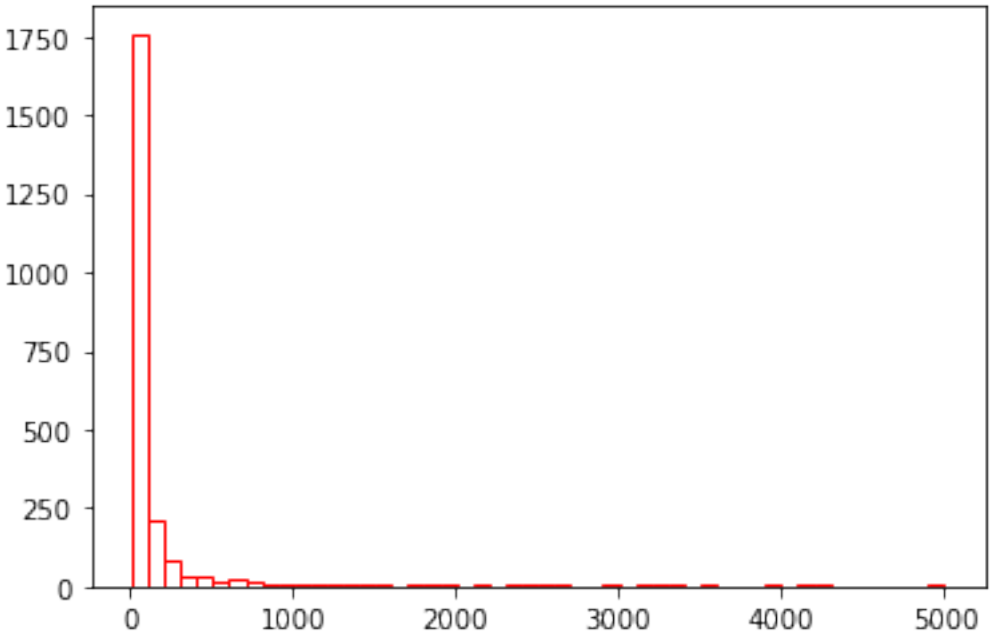
[12]:
```
fig, ax = plt.subplots(figsize = (12,12))
ax.set_title('Coeff')
cax = ax.imshow(np.sum(x_test,axis = 0), cmap = plt.cm.Accent)

cbar = plt.colorbar(cax, ticks=[0, 1, 2, 3, 4, 5, 6, 7],
                    orientation='vertical',
                    fraction=0.045, pad=0.05)

print("GREEN IS LOW, BLACK IS HIGH")
print("GREEN MIN : "+ str(int( min( np.sum(x_test,axis=0).reshape(-1,1) ) ) ))
print("BLAC MAX : "+ str(int( max( np.sum(x_test,axis=0).reshape(-1,1) ) ) ))
```

```
GREEN IS LOW, BLACK IS HIGH
GREEN MIN : 8
BLAC MAX : 4997
```



[39]:
```
print(np.sum(x_test_1d,axis=0).shape)
plt.hist(np.sum(x_test_1d,axis=0),bins=50,color='white', edgecolor='red')
plt.show()
```

```
(2304,)
```

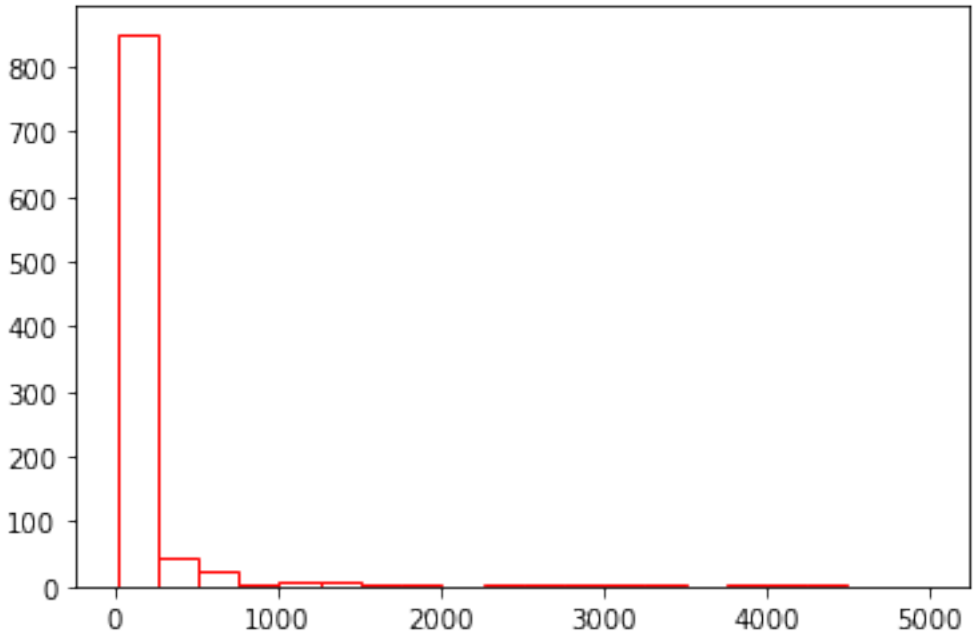


Simialr to last linear selection, most of the features are still pretty rare with very few being that common.

2.1.1 Plotting the same histogram but for MAL and BEN features

BEN FEATURES FREQ

[48]:
```
plt.hist(np.sum(x_test_1d,axis=0)[BEN_F],bins=20,color='white', edgecolor='red')
plt.show()
print(np.median(np.sum(x_test_1d,axis=0)[BEN_F]))
print(max(np.sum(x_test_1d,axis=0)[BEN_F]))
print(min(np.sum(x_test_1d,axis=0)[BEN_F]))
```




```
47.0
4997.0
12.0
```

[56]:
```
xxx = np.sum(x_test_1d,axis=0)
index = 0
max1 = max(np.sum(x_test_1d,axis=0)[BEN_F])
for i in range(x_test_1d.shape[1]):
    if xxx[i] == max1:
        index = i
print("MAX : "+features_1d[index] + " : " +str(max1))

xxxx = np.sum(x_test_1d,axis=0)
index1 = 0
min1 = min(np.sum(x_test_1d,axis=0)[BEN_F])
for i in range(x_test_1d.shape[1]):
    if xxxx[i] == min1:
        index1 = i
print("MIN : "+features_1d[index1] + " : " +str(min1))
```

```
MAX : intents::android_intent_action_MAIN : 4997.0
MIN : intents::android_intent_category_OPENABLE : 12.0
```

As we can see, though there are a few features that contribute to BEN that are super freq(as in over 4000), they are around 4500. A lot are still in the low end. But the feature with the highest freq, apearing in 4997 of the 5000 rows is a BEN feature.

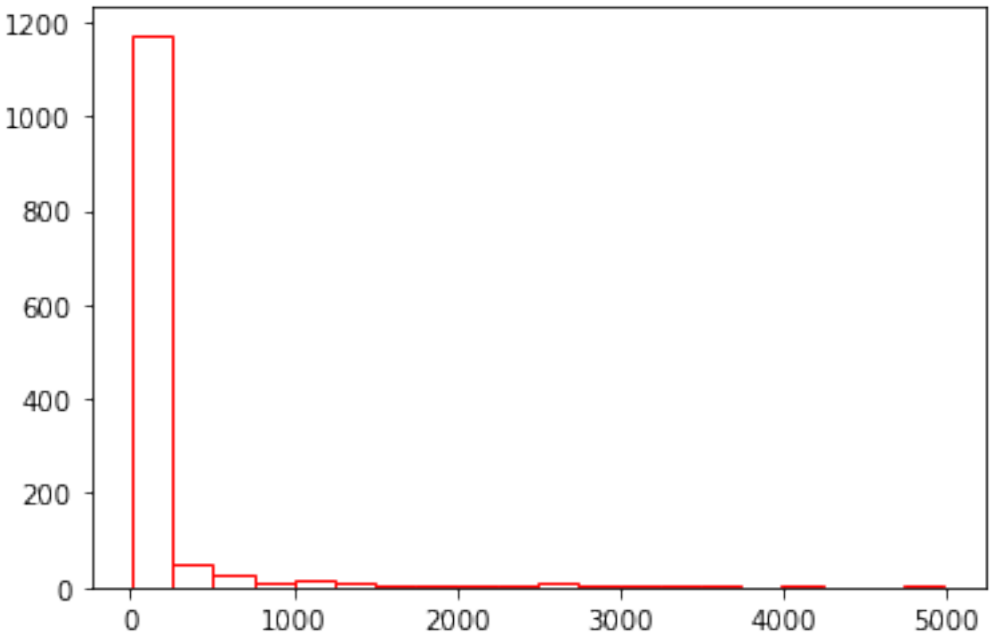
As we can see, the median freq of BEN Features is 47. We dont take mean cuz of the outrageous outliers

The most common feature, as one would expect is a default android func. ANDROID_INTENT_ACTION, appearing a whooping 4997 times

The least is intents::android_intent_category_OPENABLE that appears just 12 times

MAL FEATURES FREQ

[57]:
```
plt.hist(np.sum(x_test_1d,axis=0)[MAL_F],bins=20,color='white', edgecolor='red')
plt.show()
print(np.median(np.sum(x_test_1d,axis=0)[MAL_F]))
print(max(np.sum(x_test_1d,axis=0)[MAL_F]))
print(min(np.sum(x_test_1d,axis=0)[MAL_F]))
```




```
38.0
4989.0
8.0
```

[58]:
```
xxx = np.sum(x_test_1d,axis=0)
index = 0
max1 = max(np.sum(x_test_1d,axis=0)[MAL_F])
for i in range(x_test_1d.shape[1]):
    if xxx[i] == max1:
        index = i
print("MAX : "+features_1d[index] + " : " +str(max1))
```

```
xxxx = np.sum(x_test_1d,axis=0)
index1 = 0
min1 = min(np.sum(x_test_1d,axis=0)[MAL_F])
for i in range(x_test_1d.shape[1]):
    if xxxx[i] == min1:
        index1 = i
print("MIN : "+features_1d[index1] + " : " +str(min1))
```

```
MAX : intents::android_intent_category_LAUNCHER : 4989.0
MIN : app_permissions::name='android_permission_RECEIVE_WAP_PUSH' : 8.0
```

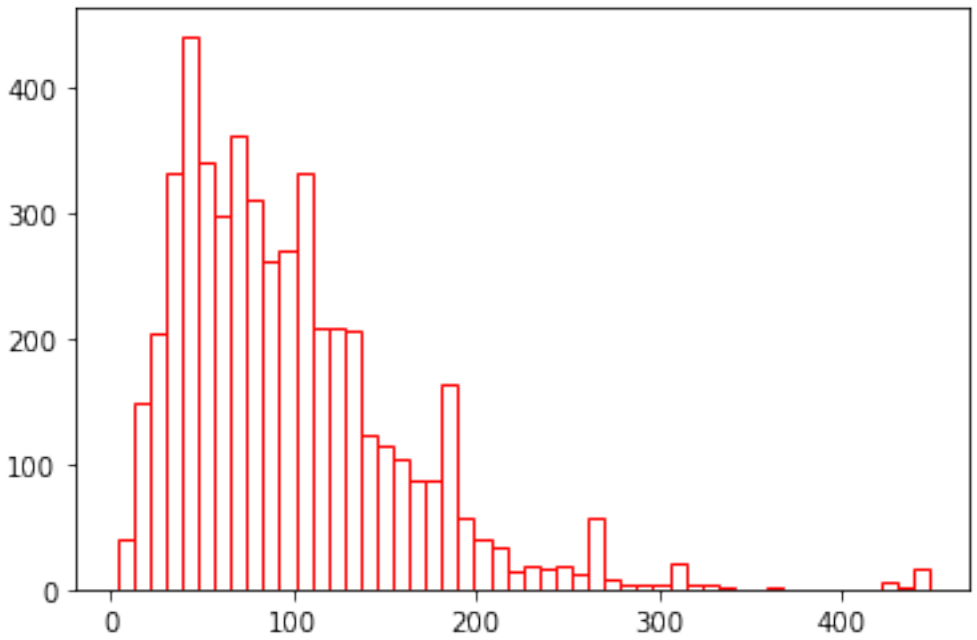
The most coomon mal feature is intents::android_intent_category_LAUNCHER that appears 4989 times. though its an android func, analysis led us to conclude that it is more often contributing towards the maliciousness.

The least common Mal feature is android_permission_RECEIVE_WAP_PUSH which appears just 8 time

2.2 Number of Features in each APK

[61]:
```
print(np.sum(x_test_1d,axis=1).shape)
plt.hist(np.sum(x_test_1d,axis=1),bins=50,color='white', edgecolor='red')
plt.show()
print(max(np.sum(x_test_1d,axis=1)))
print(min(np.sum(x_test_1d,axis=1)))
print(np.mean(np.sum(x_test_1d,axis=1)))
print(np.median(np.sum(x_test_1d,axis=1)))
```

```
(5000,)
```



```
448.0
3.0
96.5462
84.0
```

As we can see, most features have a little less than 100 features with median at 84 and mean at 96.5(even after the huge apks with over 400 features existing).

There are also apks with less than 10 features with min being 3 in our dataset. The max is 448

3 Obseravtions - FULL DATASET

- Graph of overall freq
  - Most features are rare(lots of green)
  - On the other end of the spectrum, few features appear in a lot of rows, as seen with max value of 4997 out of the 5000 rows
  - The almost neutral features are almost always rare (central area) while the features that contribute to MAL/BEN tend to be more frequent
- FOR OTHERS - INFER FROM GRAPHS

3.1 Find all MAL and BEN rows

[13]:
```
MAL = []
BEN = []
for i in range(y_test.shape[0]):
    if y_test[i]==1:
        MAL.append(i)
    else:
        BEN.append(i)
print(len(MAL),len(BEN))
```

```
1667 3333
```

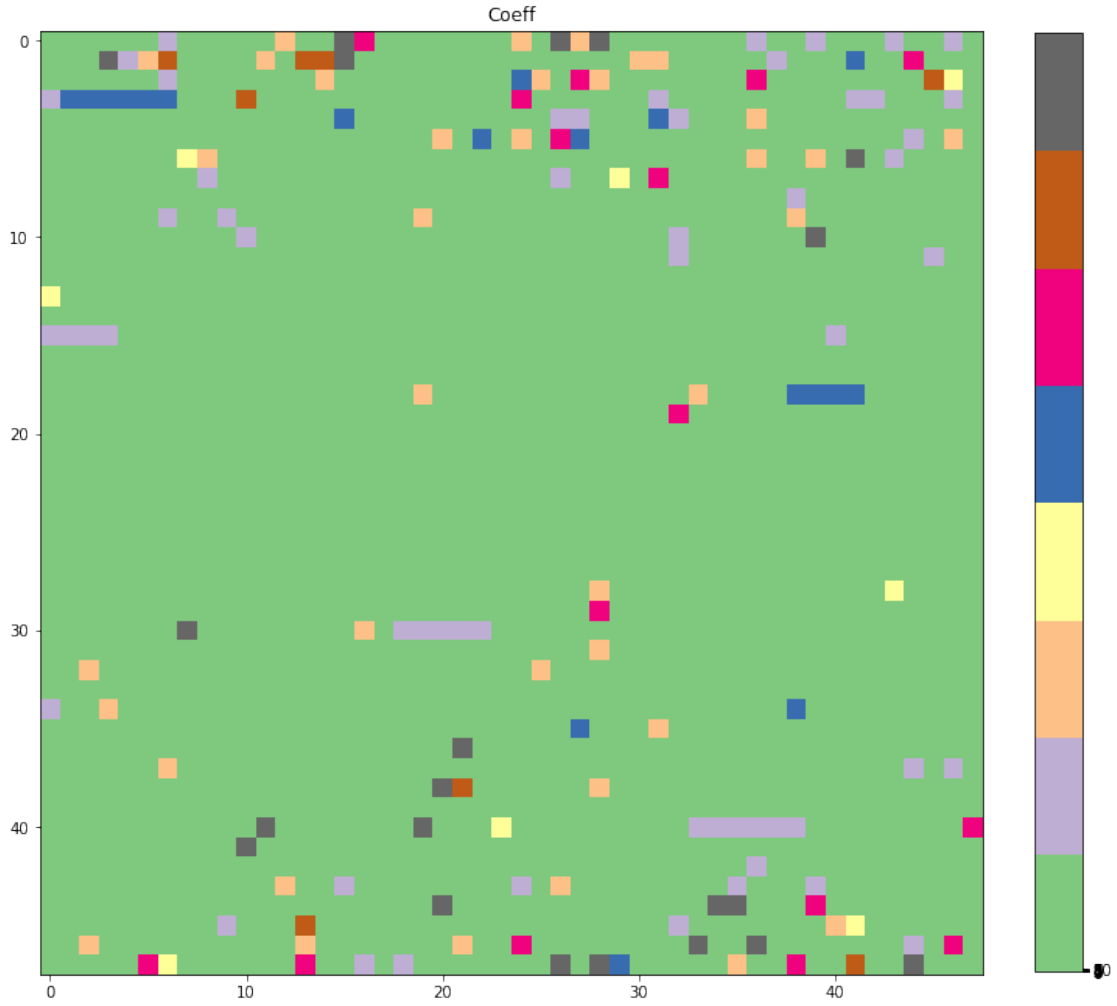
We have 1667 MAL and 3333 BEN samples

4 Seeing Feature Frequency and Counts for BEN Samples

[110]:
```
fig, ax = plt.subplots(figsize = (12,12))
ax.set_title('Coeff')
cax = ax.imshow(np.sum(x_test[BEN],axis = 0), cmap = plt.cm.Accent)

cbar = plt.colorbar(cax, ticks=[0, 1, 2, 3, 4, 5, 6, 7,8,9,10],
                    orientation='vertical',
                    fraction=0.045, pad=0.05)

print("GREEN IS LOW, BLACK IS HIGH")
```

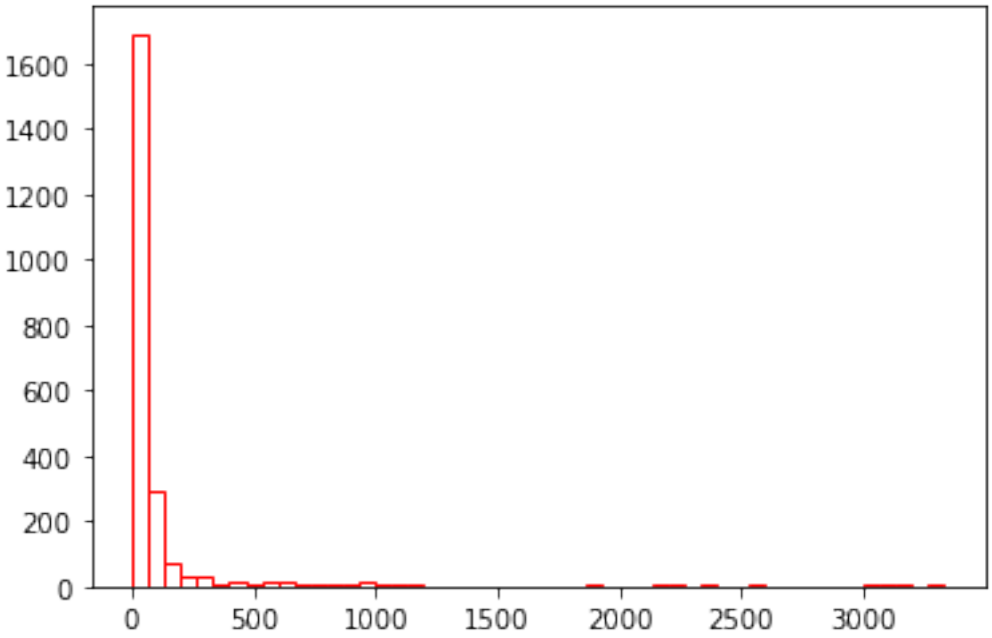
```
print("GREEN MIN : "+ str(int( min( np.sum(x_test[BEN],axis=0).reshape(-1,1) )␣
 ↪) ))
print("BLACK MAX : "+ str(int( max( np.sum(x_test[BEN],axis=0).reshape(-1,1) )␣
 ↪) ))
```

```
GREEN IS LOW, BLACK IS HIGH
GREEN MIN : 0
BLACK MAX : 3330
```



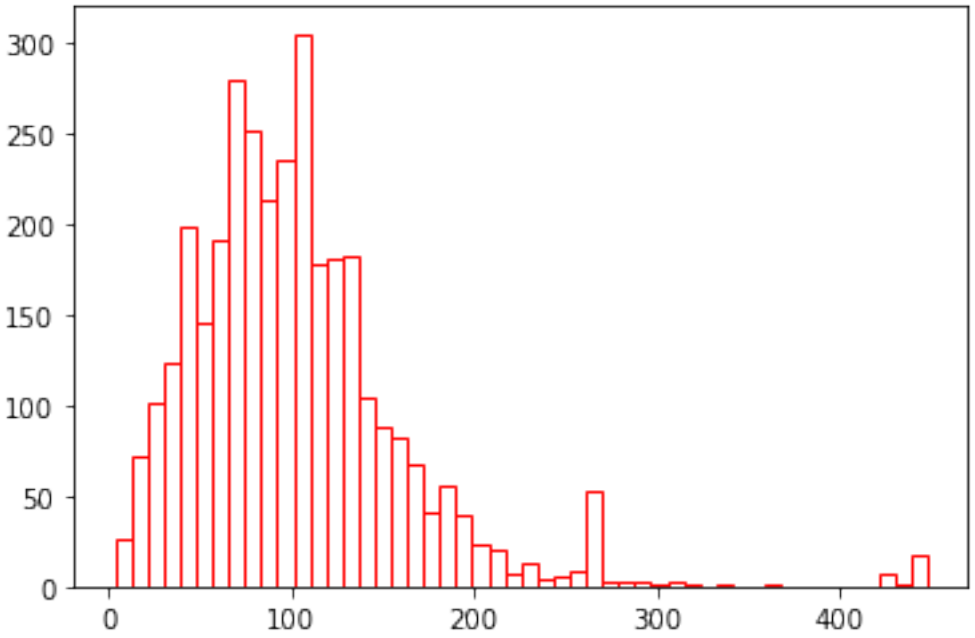
[63]:
```
print(np.sum(x_test_1d[BEN],axis=0).shape)
plt.hist(np.sum(x_test_1d[BEN],axis=0),bins=50,color='white', edgecolor='red')
plt.show()
```

```
(2304,)
```



[74]:
```
print(np.sum(x_test_1d[BEN],axis=1).shape)
plt.hist(np.sum(x_test_1d[BEN],axis=1),bins=50,color='white', edgecolor='red')
plt.show()
print("==================")
print("MAX")
print(max(np.sum(x_test_1d[BEN],axis=1)))
print("__________________")
print("MIN")
print(min(np.sum(x_test_1d[BEN],axis=1)))
print("__________________")
print("MEAN")
print(np.mean(np.sum(x_test_1d[BEN],axis=1)))
print("__________________")
print("MEDIAN")
print(np.median(np.sum(x_test_1d[BEN],axis=1)))
print("__________________")
print("====================================")
```

```
(3333,)
```




```
==================
MAX
448.0

-----------------
MIN
3.0

-----------------
MEAN
101.8904890489049

-----------------
MEDIAN
94.0

-----------------
====================================
```

[94]:
```
X_BEN_MalFets=x_test_1d[BEN][:,[MAL_F]]
```

[106]:
```
X_BEN_MalFets_SUM_BY_ROWS=np.sum(X_BEN_MalFets,axis=0)
print(np.max(X_BEN_MalFets_SUM_BY_ROWS))
print(np.min(X_BEN_MalFets_SUM_BY_ROWS))
print(np.mean(X_BEN_MalFets_SUM_BY_ROWS))
print(np.median(X_BEN_MalFets_SUM_BY_ROWS))
```

```
3326.0
0.0
137.56132430398796
```

```
25.0
```

[91]:
```
X_BEN_BenFets=x_test_1d[BEN][:,[BEN_F]]
```

[107]:
```
X_BEN_BenFets_SUM_BY_ROWS=np.sum(X_BEN_BenFets,axis=0)
print(np.max(X_BEN_BenFets_SUM_BY_ROWS))
print(np.min(X_BEN_BenFets_SUM_BY_ROWS))
print(np.mean(X_BEN_BenFets_SUM_BY_ROWS))
print(np.median(X_BEN_BenFets_SUM_BY_ROWS))
```

```
3330.0
0.0
160.80205128205128
29.0
```

4.1 Obseravtions - BEN Rows

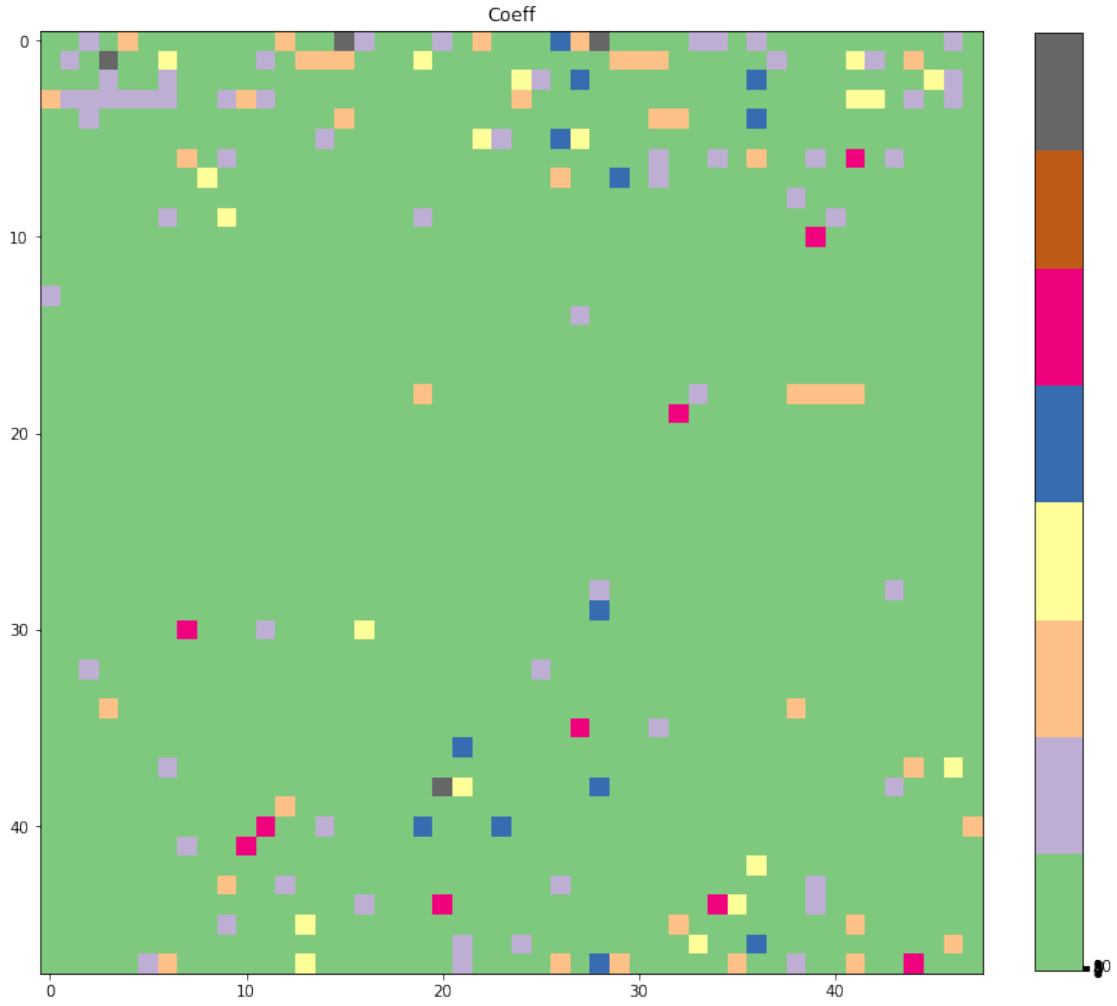
- Considering just BEN Rows
  - A BEN Feature appears a median of 29 times
  - A MAL Feature appears a median of 25 times
  - A BEN APK has a median of 94 features of which more than half are usually BEN
  - We just have 3333 BEn out of which there is a feature with 3330 appearances. On the other hand, there do exist features with 0 appearances in BEN, these features are not necessarily just BEN and are MAL fets and are also sometimes BEN_Fs

[ ]:
```
```

5 Seeing Feature Frequency and Counts for MAL Samples

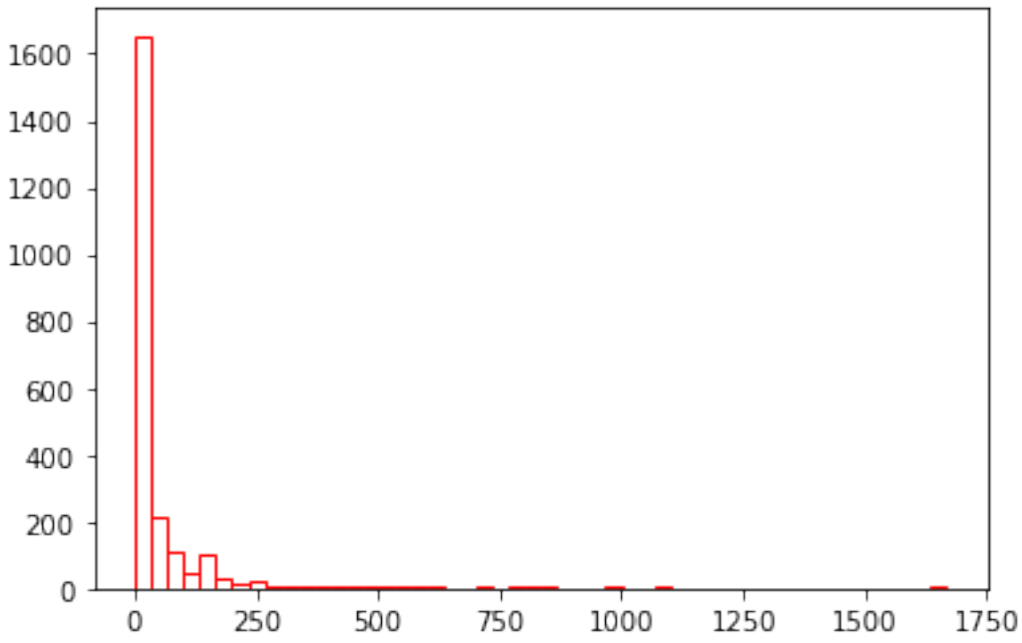
[111]:
```
fig, ax = plt.subplots(figsize = (12,12))
ax.set_title('Coeff')
cax = ax.imshow(np.sum(x_test[MAL],axis = 0), cmap = plt.cm.Accent)

cbar = plt.colorbar(cax, ticks=[0, 1, 2, 3, 4, 5, 6, 7,8,9,10],
                    orientation='vertical',
                    fraction=0.045, pad=0.05)

print("GREEN IS LOW, BLACK IS HIGH")
print("GREEN MIN : "+ str(int( min( np.sum(x_test[MAL],axis=0).reshape(-1,1) )␣
 ↪) ))
print("BLACK MAX : "+ str(int( max( np.sum(x_test[MAL],axis=0).reshape(-1,1) )␣
 ↪) ))
```

```
GREEN IS LOW, BLACK IS HIGH
GREEN MIN : 0
BLACK MAX : 1667
```



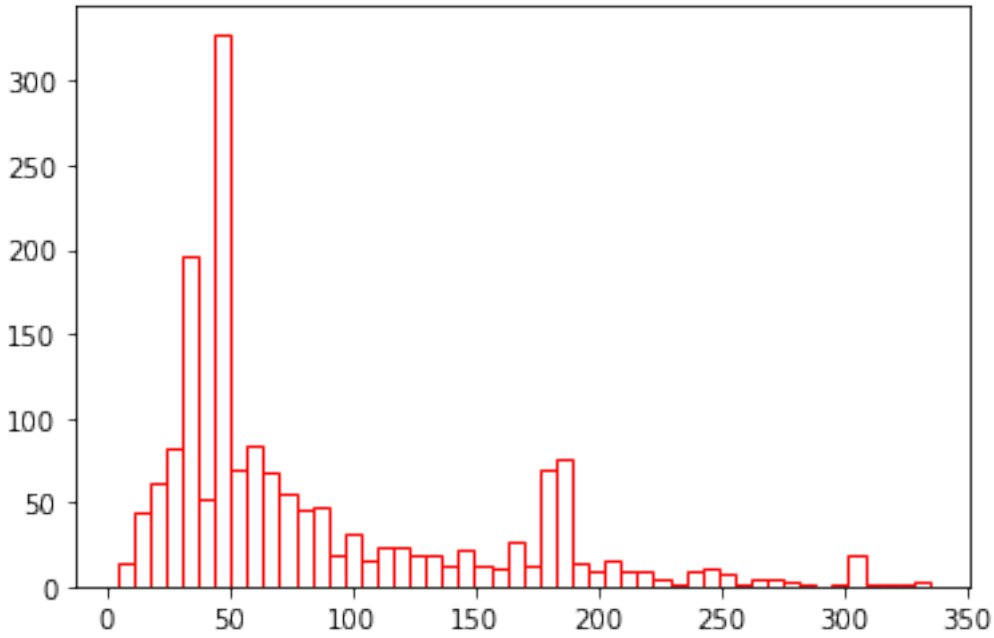
[112]:
```
print(np.sum(x_test_1d[MAL],axis=0).shape)
plt.hist(np.sum(x_test_1d[MAL],axis=0),bins=50,color='white', edgecolor='red')
plt.show()
```

```
(2304,)
```



[113]:
```
print(np.sum(x_test_1d[MAL],axis=1).shape)
plt.hist(np.sum(x_test_1d[MAL],axis=1),bins=50,color='white', edgecolor='red')
plt.show()
print("==================")
print("MAX")
print(max(np.sum(x_test_1d[MAL],axis=1)))
print("__________________")
print("MIN")
print(min(np.sum(x_test_1d[MAL],axis=1)))
print("__________________")
print("MEAN")
print(np.mean(np.sum(x_test_1d[MAL],axis=1)))
print("__________________")
print("MEDIAN")
print(np.median(np.sum(x_test_1d[MAL],axis=1)))
print("__________________")
print("====================================")
```

```
(1667,)
```




```
==================
MAX
335.0

-----------------
MIN
4.0

-----------------
MEAN
85.86082783443311

-----------------
MEDIAN
56.0

-----------------
====================================
```

[114]:
```
X_MAL_MalFets=x_test_1d[MAL][:,[MAL_F]]
```

[115]:
```
X_MAL_MalFets_SUM_BY_ROWS=np.sum(X_MAL_MalFets,axis=0)
print(np.max(X_MAL_MalFets_SUM_BY_ROWS))
print(np.min(X_MAL_MalFets_SUM_BY_ROWS))
print(np.mean(X_MAL_MalFets_SUM_BY_ROWS))
print(np.median(X_MAL_MalFets_SUM_BY_ROWS))
```

```
1663.0
0.0
62.2686230248307
```

```
17.0
```

[116]:
```
X_MAL_BenFets=x_test_1d[MAL][:,[BEN_F]]
```

[117]:
```
X_MAL_BenFets_SUM_BY_ROWS=np.sum(X_MAL_BenFets,axis=0)
print(np.max(X_MAL_BenFets_SUM_BY_ROWS))
print(np.min(X_MAL_BenFets_SUM_BY_ROWS))
print(np.mean(X_MAL_BenFets_SUM_BY_ROWS))
print(np.median(X_MAL_BenFets_SUM_BY_ROWS))
```

```
1667.0
0.0
61.92307692307692
12.0
```

5.1 Obseravtions - MAL Rows

- Considering just MAL Rows
  - A BEN Feature appears a median of 12 times
  - A MAL Feature appears a median of 17 times
  - There numbers might so thats not a huge concern given the dataset. After all, we have 2304 features and just 1667 MAL APK Rows
  - A BEN APK has a median of 56 features of which the majority is usually MAL
  - There are features that appear 0 times in MAL APKs and also ones that appear in all rows

[ ]:
```
```