Iterative Clustering_1st Iteration

April 14, 2020

1 Image Clustering

```
Using TensorFlow backend.
```

```
Total number of files  1532
Total number of columns 3
```

1.1 Classwise Data Distribution

```
[6]:             ClassName  NumFiles
     0                 PAN       179
     1      Medical Report        12
     2           Form 1040       100
     3           Insurance        36
     4   Airtel Mobile Bill       43
     5           Form 6251       100
     6            Passport       192
     7      Bank Statement         9
     8   NewgenVisitingCard      159
     9              Resume       120
     10          Form 2441       100
     11         NewgenIDs        108
     12   Electricity Bill        32
     13          Form 2106       100
     14   Credit Card Bills        18
     15            DubaiID        86
     16         Floor Plan        27
     17             Aadhar        91
     18                IGL        20
```

2 Minimum threshold = 50 images

- We will remove those classes which fail to pass the condition that total number of files in that class<50

```
Classes having images count less than 50 -->
Medical Report
Insurance
Airtel Mobile Bill
Bank Statement
Electricity Bill
Credit Card Bills
Floor Plan
IGL

The updated number of data files left 1335
The number of classes to be considered for first iteration 11
```

3 Cluster Data Distribution Matrix

[36]:

```
                      0   1    2    3   4    5    6    7    8   9   10
DubaiID               0   0    0    0  60   12    0    0    0  14    0
Form 6251             0   0    0    0   0    0    0    0    0   0  100
Resume                0   0    0    0   0    0    0    0  120   0    0
NewgenIDs             0   0    0    0   0    0    0  108    0   0    0
Form 2106             0   0    0    0   0    0  100    0    0   0    0
Aadhar                0   0    0    0   0    0    0    0    0  91    0
Form 2441             0   0    0    0   0    0    0    0    0   0  100
Form 1040           100   0    0    0   0    0    0    0    0   0    0
NewgenVisitingCard    0   0    1  156   0    0    0    0    0   2    0
Passport              0   0  151    0   0    0    0    0    0  41    0
PAN                   0  59    0    0   0  120    0    0    0   0    0
```

```
KMeans VGG19 (PCA):
```

4 RELATIVE PURITY MATRIX

Relative purity of a cluster is defined as

relative purity of cluster 1 w.r.t class_a = (number of samples of class_a in the cluster 1/total number of samples in cluster 1)*100

which implies that 100% value indicates that the cluster does not contain samples from any other class

We set the benchmark for purity of a cluster a **relative percentage to be minimum 90 for first iteration** to be identified as a pure cluster

[40]:

```
                      0      1     2      3      4     5      6      7  \
DubaiID             0.0    0.0   0.0    0.0  100.0   9.0    0.0    0.0
Form 6251           0.0    0.0   0.0    0.0    0.0   0.0    0.0    0.0
Resume              0.0    0.0   0.0    0.0    0.0   0.0    0.0    0.0
```

```
NewgenIDs             0.0    0.0   0.0    0.0    0.0   0.0    0.0  100.0
Form 2106             0.0    0.0   0.0    0.0    0.0   0.0  100.0    0.0
Aadhar                0.0    0.0   0.0    0.0    0.0   0.0    0.0    0.0
Form 2441             0.0    0.0   0.0    0.0    0.0   0.0    0.0    0.0
Form 1040           100.0    0.0   0.0    0.0    0.0   0.0    0.0    0.0
NewgenVisitingCard    0.0    0.0   1.0  100.0    0.0   0.0    0.0    0.0
Passport              0.0    0.0  99.0    0.0    0.0   0.0    0.0    0.0
PAN                   0.0  100.0   0.0    0.0    0.0  93.0    0.0    0.0

                        8     9    10
DubaiID               0.0   9.0   0.0
Form 6251             0.0   0.0  50.0
Resume              100.0   0.0   0.0
NewgenIDs             0.0   0.0   0.0
Form 2106             0.0   0.0   0.0
Aadhar                0.0  64.0   0.0
Form 2441             0.0   0.0  67.0
Form 1040             0.0   0.0   0.0
NewgenVisitingCard    0.0   2.0   0.0
Passport              0.0  35.0   0.0
PAN                   0.0   0.0   0.0
```

5 PURE CLUSTERS ITERATION-1

[42]:

```
             ClassName  ClusterNumber  PurityPercentage  ClassPercentage
0              DubaiID              4             100.0             70.0
1               Resume              8             100.0            100.0
2            NewgenIDs              7             100.0            100.0
3            Form 2106              6             100.0            100.0
4            Form 1040              0             100.0            100.0
5   NewgenVisitingCard              3             100.0             98.0
6             Passport              2              99.0             79.0
7                  PAN              1             100.0             33.0
8                  PAN              5              93.0             67.0
```

6 OBSERVATION

The above tabel contains the pure clusters with the class percentage

where class percentage is defined as

class percentage = (number of samples of class in that particular cluster)/(total number of samples of that class)

We observe that we find 13 pure clusters out of 18 initial clusters initially formed

7 Finding the Correctly identified classes

we create a benchmark for correctly identified classes as **classpercentage > 80.0%**

First we will club all the pure clusters of the classes

if their sum >80.0 then the class is most suitable for image based clustering

7.1 Pure Class Report-1 Iteration

[56]:

```
            className  ClassPercentage  CorrectlyClassifiedFiles  TotalFiles
0             DubaiID             70.0                        60          86
1              Resume            100.0                       120         120
2           NewgenIDs            100.0                       108         108
3           Form 2106            100.0                       100         100
4           Form 1040            100.0                       100         100
5  NewgenVisitingCard             98.0                       155         159
6            Passport             79.0                       151         192
7                 PAN            100.0                       179         179
```

```
The correctly identified classes in first iterations are--->
Resume
NewgenIDs
Form 2106
Form 1040
NewgenVisitingCard
PAN
```

Iteration-2 Clustering

April 14, 2020

1 Second Iteration Clustering

```
Using TensorFlow backend.
```

1.1 Class Wise Data Distribution

```
[9]:     ClassName  NumberofFiles
     0  Form 6251            100
     1  Form 2441            100
     2    DubaiID             14
     3   Passport             41
     4     Aadhar             91
```

```
The number of clusters for second iteration 10
Number of Rows  346
```

2 DATA DISTRIBUTION MATRIX

```
KMeans VGG19:
```

```
[26]:             0    1    2   3   4   5  6   7   8   9
      Form 6251   0    0  100   0   0   0  0   0   0   0
      Form 2441   0  100    0   0   0   0  0   0   0   0
      DubaiID     0    0    0   3   0   0  8   0   3   0
      Passport    0    0    0   0   0  38  1   0   2   0
      Aadhar     17    0    0  15  12   0  8  16  13  10
```

3 RELATIVE PURITY MATRIX

```
[29]:               0      1      2     3      4      5     6      7     8      9
      Form 6251   0.0    0.0  100.0   0.0    0.0    0.0   0.0    0.0   0.0    0.0
      Form 2441   0.0  100.0    0.0   0.0    0.0    0.0   0.0    0.0   0.0    0.0
      DubaiID     0.0    0.0    0.0  17.0    0.0    0.0  47.0    0.0  17.0    0.0
      Passport    0.0    0.0    0.0   0.0    0.0  100.0   2.0    0.0   6.0    0.0
      Aadhar     100.0     0.0     0.0  47.0  100.0     0.0  14.0  100.0  36.0  100.0
```

3.1 Clusters of Pure Classes

```
[30]:    ClassName  ClusterNumber  PurityPercentage  ClassPercentage
      0  Form 6251              2             100.0            100.0
      1  Form 2441              1             100.0            100.0
      2   Passport              5             100.0             93.0
      3     Aadhar              0             100.0             19.0
      4     Aadhar              4             100.0             13.0
      5     Aadhar              7             100.0             18.0
      6     Aadhar              9             100.0             11.0
```

3.2 Pure Class Report-2 Iteration

```
[33]:    className  ClassPercentage  CorrectlyClassifiedFiles  TotalFiles
      0  Form 6251            100.0                       100         100
      1  Form 2441            100.0                       100         100
      2   Passport             93.0                        38          41
      3     Aadhar             61.0                        55          91
```

3.3 Pure Class Report-1 Iteration

```
[34]:             className  ClassPercentage  CorrectlyClassifiedFiles  TotalFiles
      0             DubaiID             70.0                        60          86
      1              Resume            100.0                       120         120
      2           NewgenIDs            100.0                       108         108
      3           Form 2106            100.0                       100         100
      4           Form 1040            100.0                       100         100
      5  NewgenVisitingCard             98.0                       155         159
      6            Passport             79.0                       151         192
      7                 PAN            100.0                       179         179
```

3.4 Combined Report

```
[36]:              className  ClassPercentage_x  CorrectlyClassifiedFiles_x  \
      0              DubaiID               70.0                        60.0
      1               Resume              100.0                       120.0
      2            NewgenIDs              100.0                       108.0
      3            Form 2106              100.0                       100.0
      4            Form 1040              100.0                       100.0
      5   NewgenVisitingCard               98.0                       155.0
```

```
6             Passport               79.0                       151.0
7                  PAN              100.0                       179.0
8            Form 6251                NaN                         NaN
9            Form 2441                NaN                         NaN
10              Aadhar                NaN                         NaN

    TotalFiles_x  ClassPercentage_y  CorrectlyClassifiedFiles_y  TotalFiles_y
0           86.0                NaN                         NaN           NaN
1          120.0                NaN                         NaN           NaN
2          108.0                NaN                         NaN           NaN
3          100.0                NaN                         NaN           NaN
4          100.0                NaN                         NaN           NaN
5          159.0                NaN                         NaN           NaN
6          192.0               93.0                        38.0          41.0
7          179.0                NaN                         NaN           NaN
8            NaN              100.0                       100.0         100.0
9            NaN              100.0                       100.0         100.0
10           NaN               61.0                        55.0          91.0
```

Pre Training Data Analysis

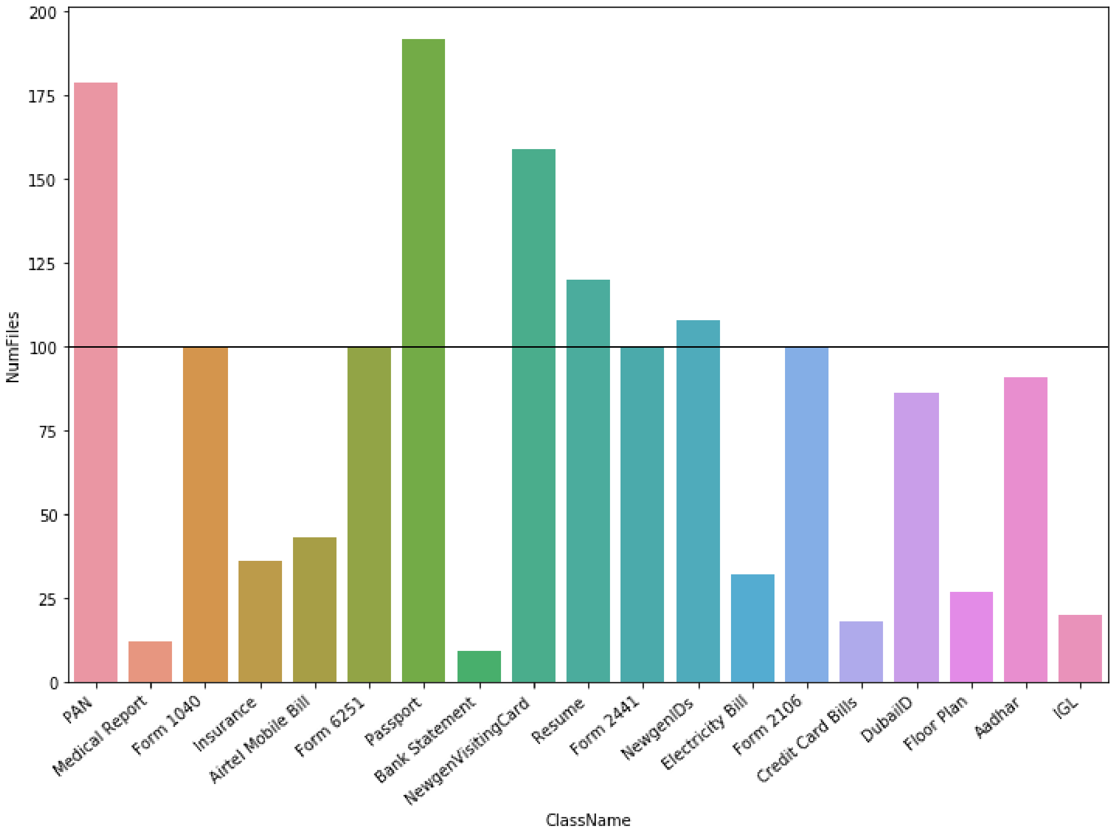
April 14, 2020

1 CLASSWISE DISTRIBUTION

```
[4]:             ClassName  NumFiles
     0                  PAN       179
     1       Medical Report        12
     2            Form 1040       100
     3            Insurance        36
     4   Airtel Mobile Bill        43
     5            Form 6251       100
     6             Passport       192
     7       Bank Statement         9
     8   NewgenVisitingCard       159
     9               Resume       120
     10           Form 2441       100
     11           NewgenIDs       108
     12    Electricity Bill        32
     13           Form 2106       100
     14   Credit Card Bills        18
     15             DubaiID        86
     16          Floor Plan        27
     17              Aadhar        91
     18                 IGL        20
```

1.1 Bar plot with threshold line

We check for all classes which don’t have total number of images< threshold count i.e 100



2 IMBALANCED DATASET FACTOR

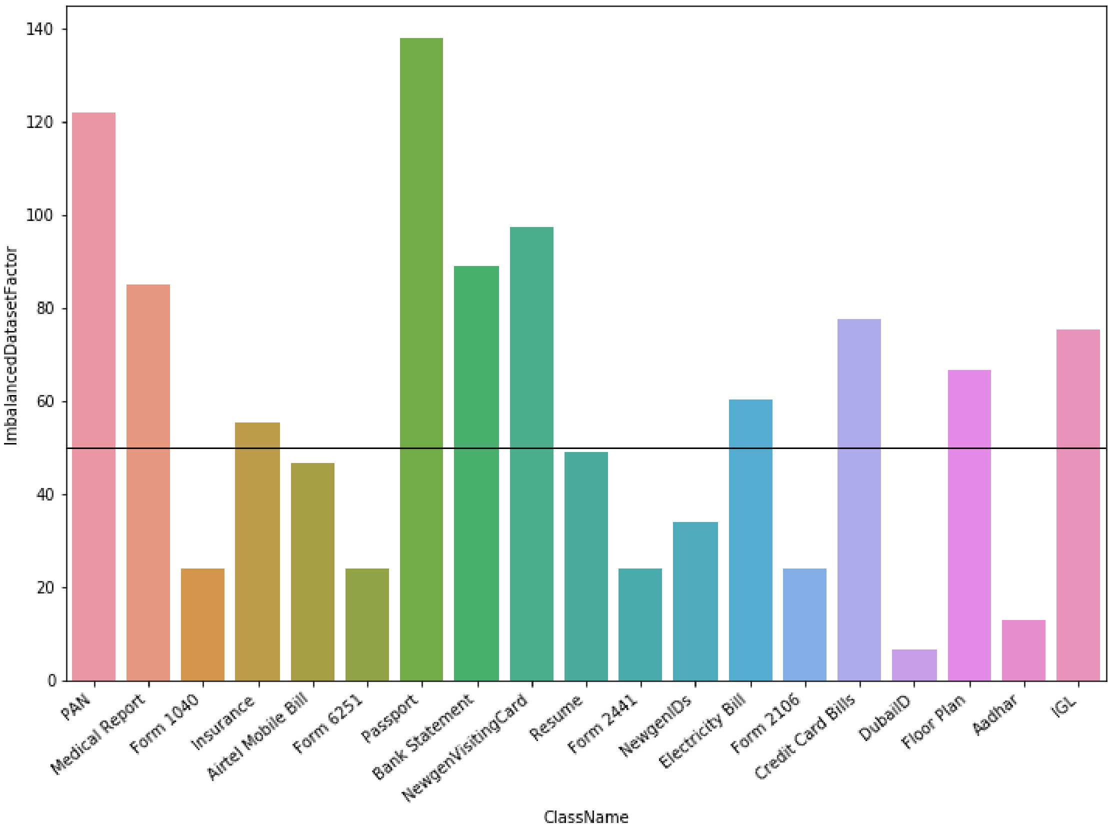
We define **imbalanced dataset factor** = |Number of files(for that class) - Average number of files|/Average number of files)*100

```
Average Number of files per class 80.63157894736842
```

```
[6]:             ClassName  NumFiles  ImbalancedDatasetFactor
     0                 PAN       179                   122.00
     1      Medical Report        12                    85.12
     2           Form 1040       100                    24.02
     3           Insurance        36                    55.35
     4  Airtel Mobile Bill        43                    46.67
     5           Form 6251       100                    24.02
     6            Passport       192                   138.12
     7      Bank Statement         9                    88.84
     8  NewgenVisitingCard       159                    97.19
     9              Resume       120                    48.83
     10          Form 2441       100                    24.02
     11          NewgenIDs       108                    33.94
```

```
12    Electricity Bill       32                     60.31
13           Form 2106      100                     24.02
14   Credit Card Bills       18                     77.68
15             DubaiID       86                      6.66
16          Floor Plan       27                     66.51
17              Aadhar       91                     12.86
18                 IGL       20                     75.20
```

2.1 Bar plot with threshold line



3 Folder statistics

The various metrics of the dataset folder are saved in the file **data_stats.csv**

[9]:

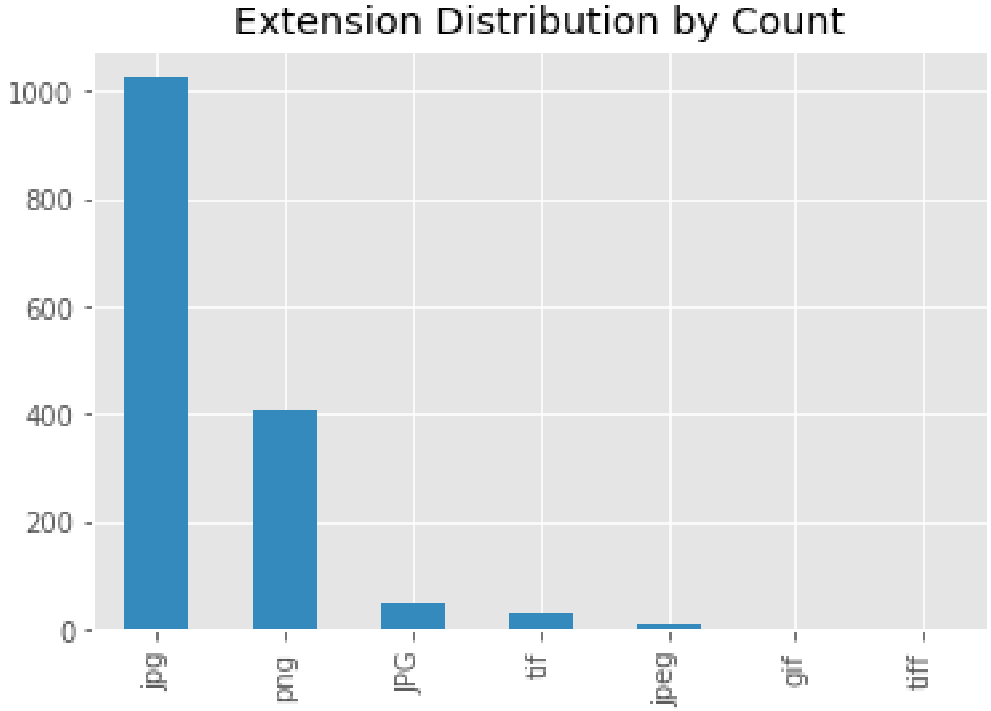
```
   id                                              path  \
0   3  /home/abhinav/dataset_analysis/image_clusterin…
1   4  /home/abhinav/dataset_analysis/image_clusterin…
2   5  /home/abhinav/dataset_analysis/image_clusterin…
```

```
3   6  /home/abhinav/dataset_analysis/image_clusterin…
4   7  /home/abhinav/dataset_analysis/image_clusterin…

                        name extension     size               atime  \
0  IMG_20181003_181804_aug_3       jpg  3977147 2020-04-14 18:19:02
1        IMG_20181011_121434       jpg  2740842 2020-04-14 18:19:02
2          PAN-CARD-ABHISHEK       jpg   275946 2020-04-14 18:19:02
3                       01_0       jpg    26360 2020-04-14 18:19:02
4             IMG_3900_aug_2       jpg  2706662 2020-04-14 18:19:02

                mtime               ctime  folder  num_files  depth  parent  \
0 2018-10-04 11:31:06 2020-04-07 11:24:58   False        NaN      1       2
1 2018-10-11 12:26:54 2020-04-07 11:25:06   False        NaN      1       2
2 2018-10-03 15:49:32 2020-04-07 11:25:09   False        NaN      1       2
3 2018-08-16 11:37:26 2020-04-07 11:24:56   False        NaN      1       2
4 2018-10-11 14:31:32 2020-04-07 11:25:08   False        NaN      1       2

    uid
0  1000
1  1000
2  1000
3  1000
4  1000
```

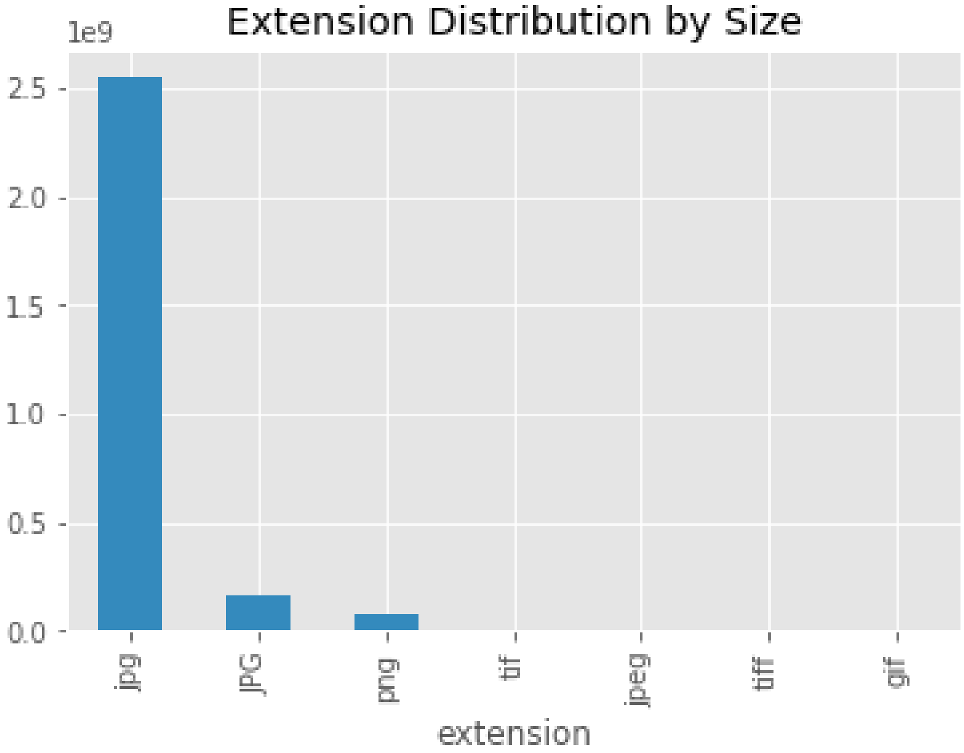
1. We get the following stats for the file in the dataset –> extension,file size, atime,mtime,ctime,checking folder,'num_files' present if that is a folder,folder depth

atime : time of last access
mtime : time of last modification
ctime : time of last status (metadata) change like file permissions, file ownership, etc. (creation time in Windows)

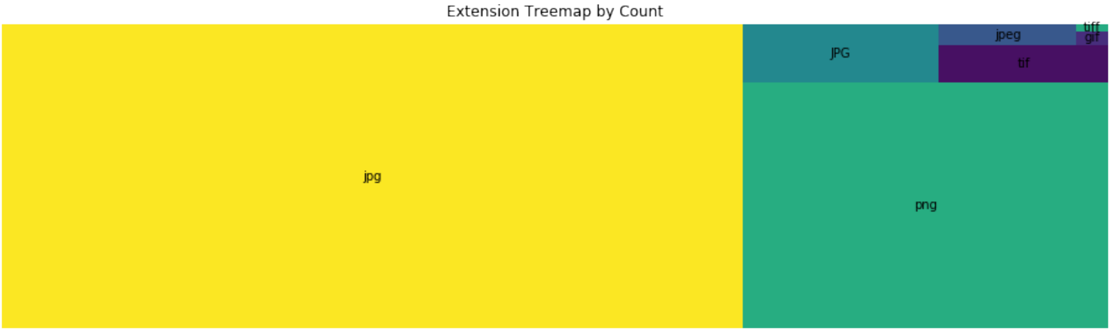
3.1 Extension distribution by count–Overall



3.2 Extension distribution by size–>Overall



3.3 Extension Treemap by Count

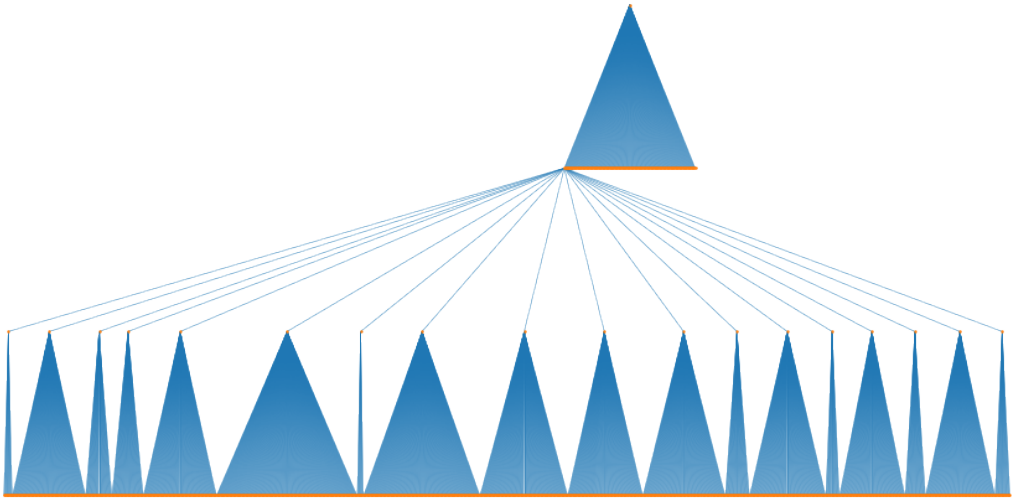


4 TREE GRAPH OF FOLDER STRUCTURE

It helps in understanding the presence of the folder structure and the presence various subfolders,depth of the dataset folder

It should be balanced for a proper dataset folder

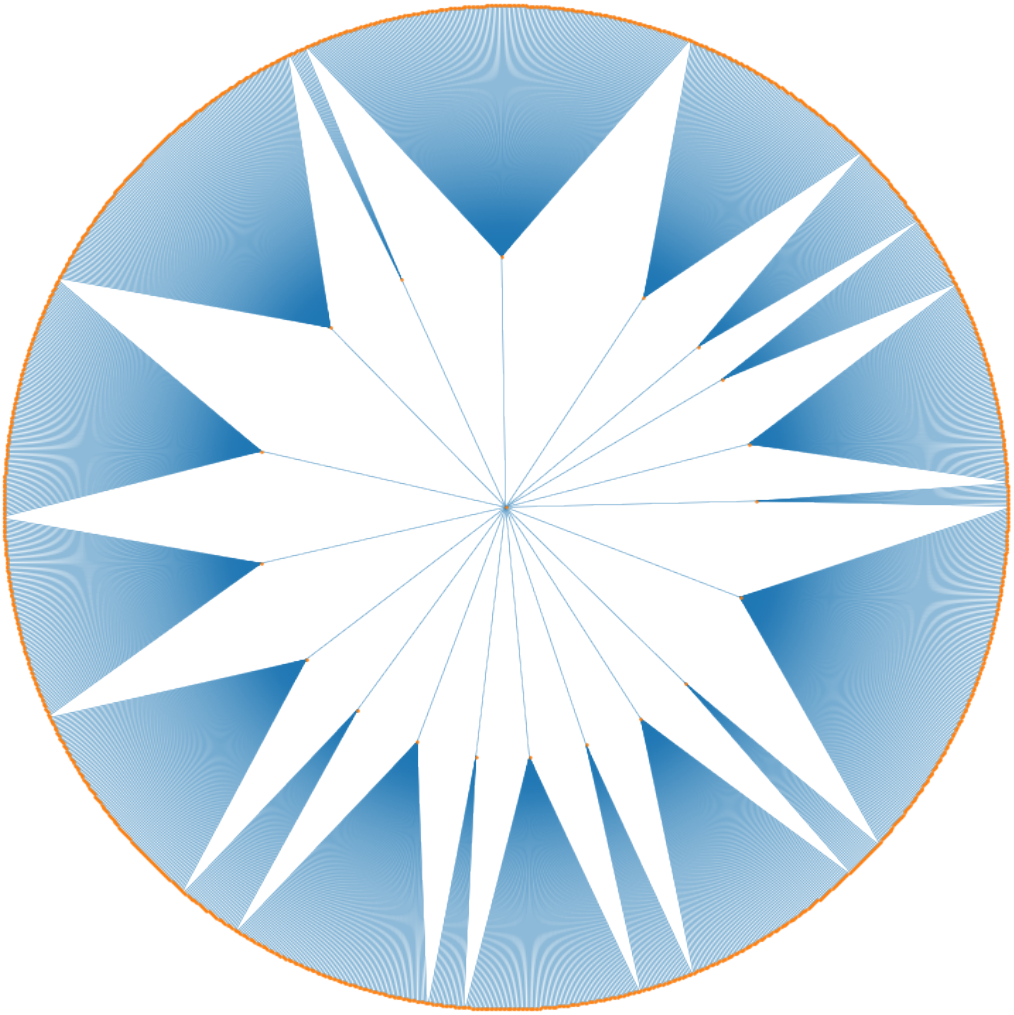
```
Name:
Type: Graph
Number of nodes: 1552
Number of edges: 1551
Average degree:   1.9987
```



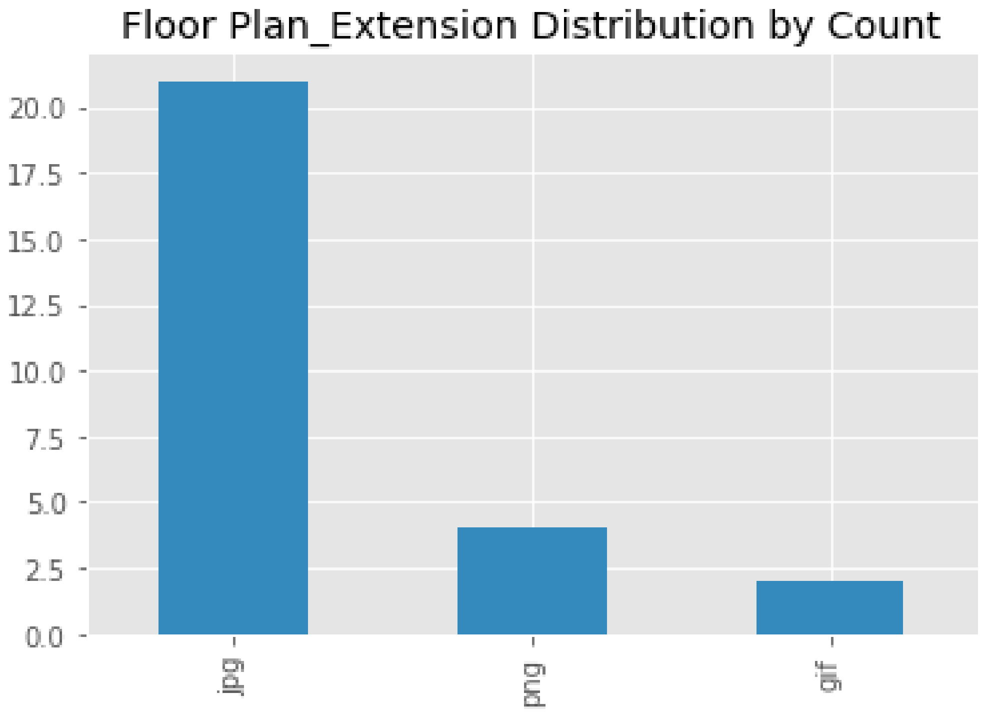
1. Number of nodes
RootNode–>Data(1)+ Total Number of classes(19)+Total Files(1532)
Hence Number of nodes = 1552
2. Number of edges
Total Number of classes(19) + Total Files(1532)
1551

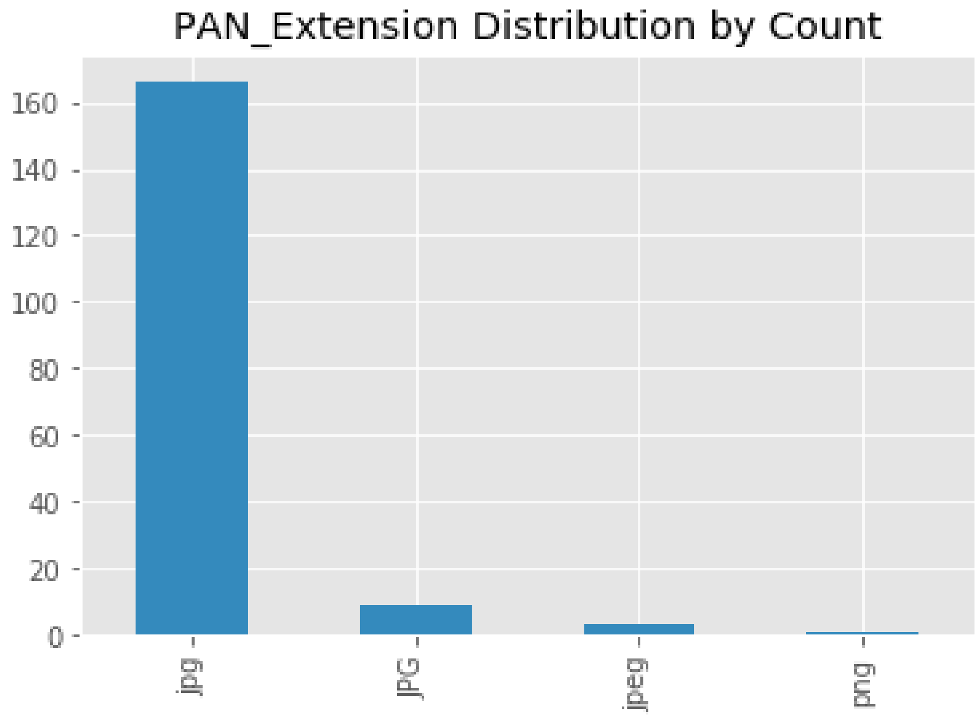
5 RADIAL GRAPH OF FOLDER STRUCTURE

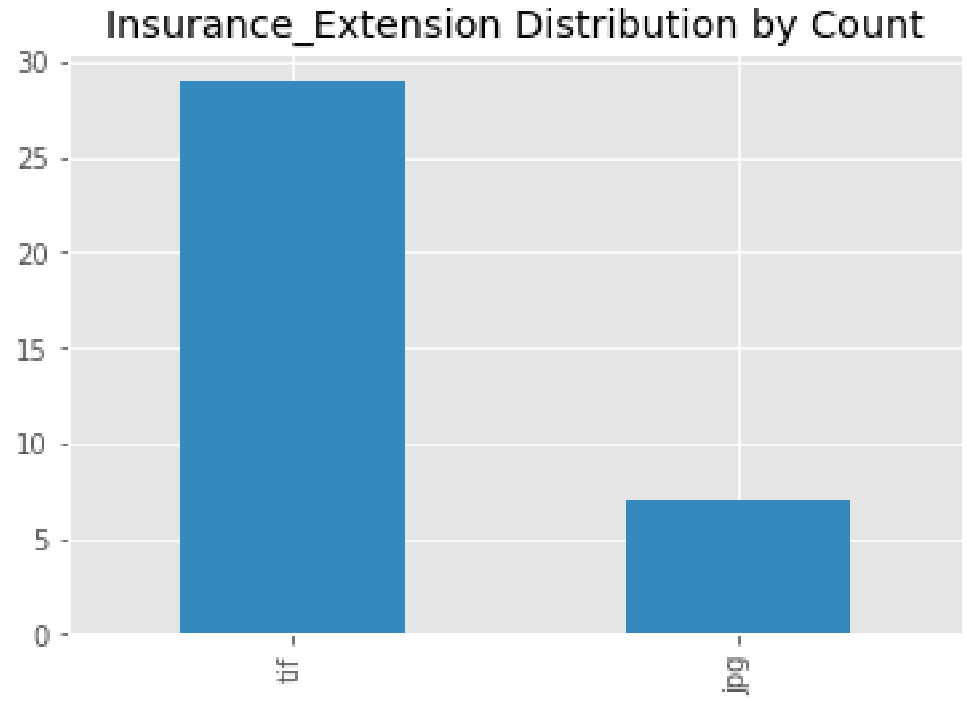
1. Large clusters represent large folder size and vice versa
2. All clusters enclosed in the circle implies that there are no subfolders

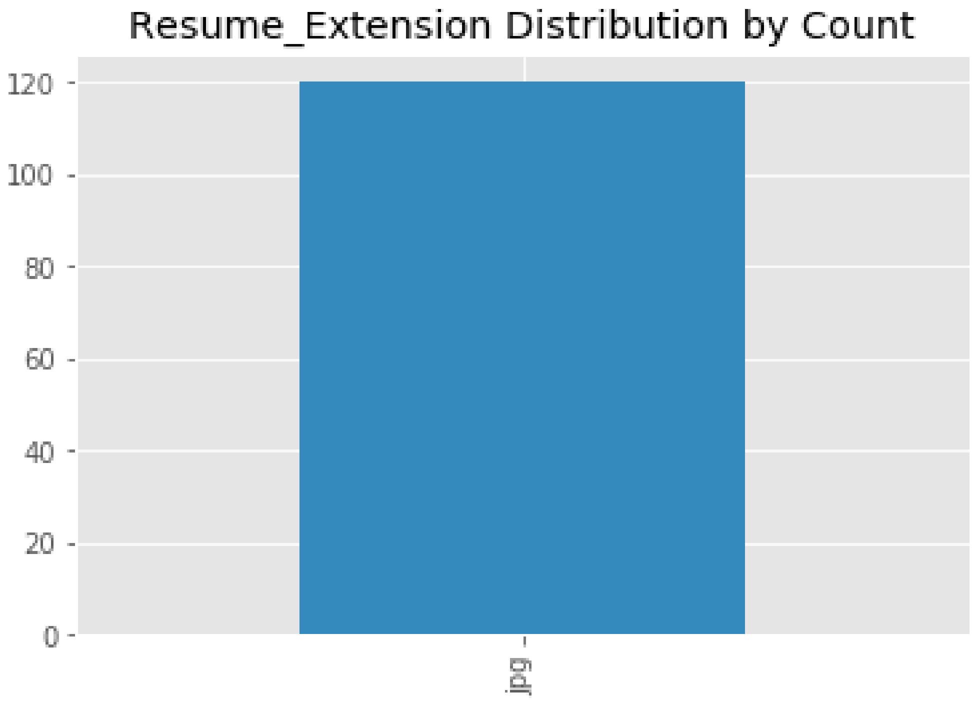


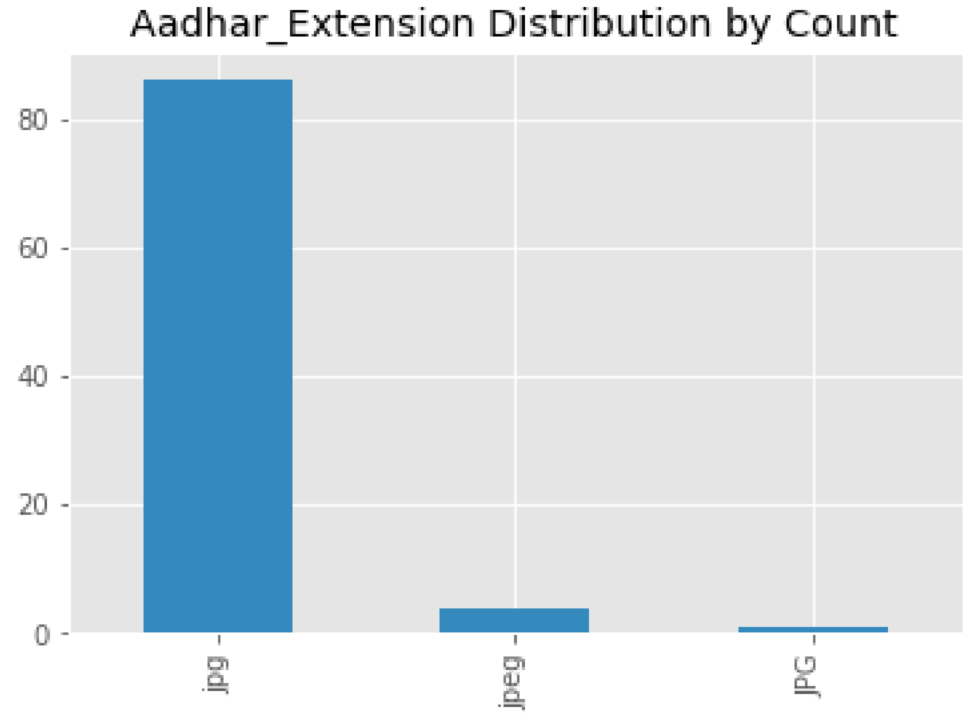
6 FORMAT WISE ANALYSIS PER CLASS

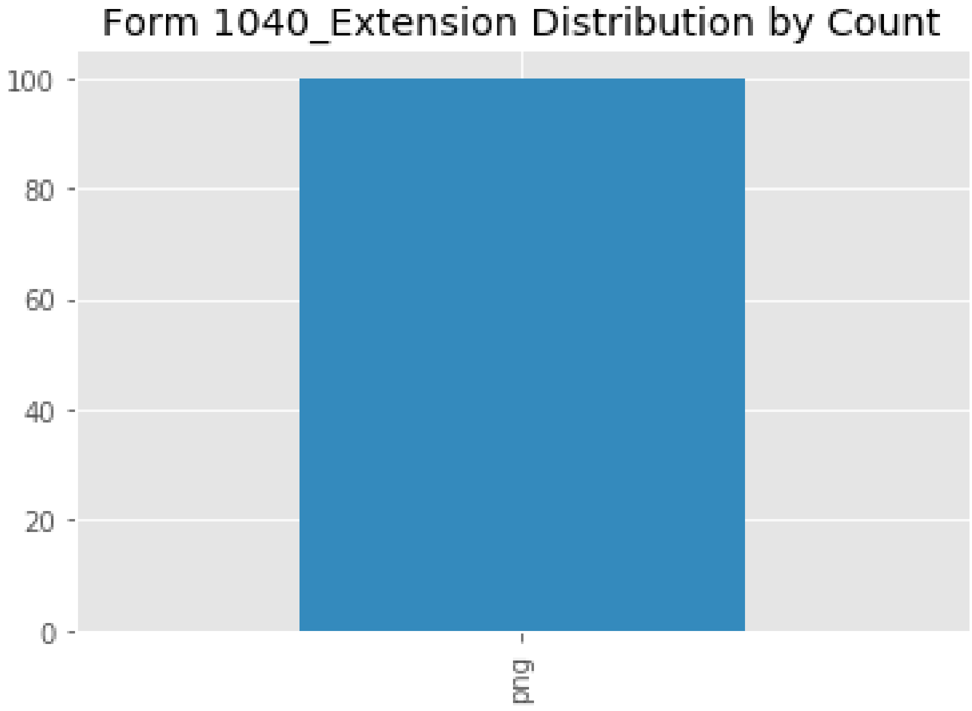


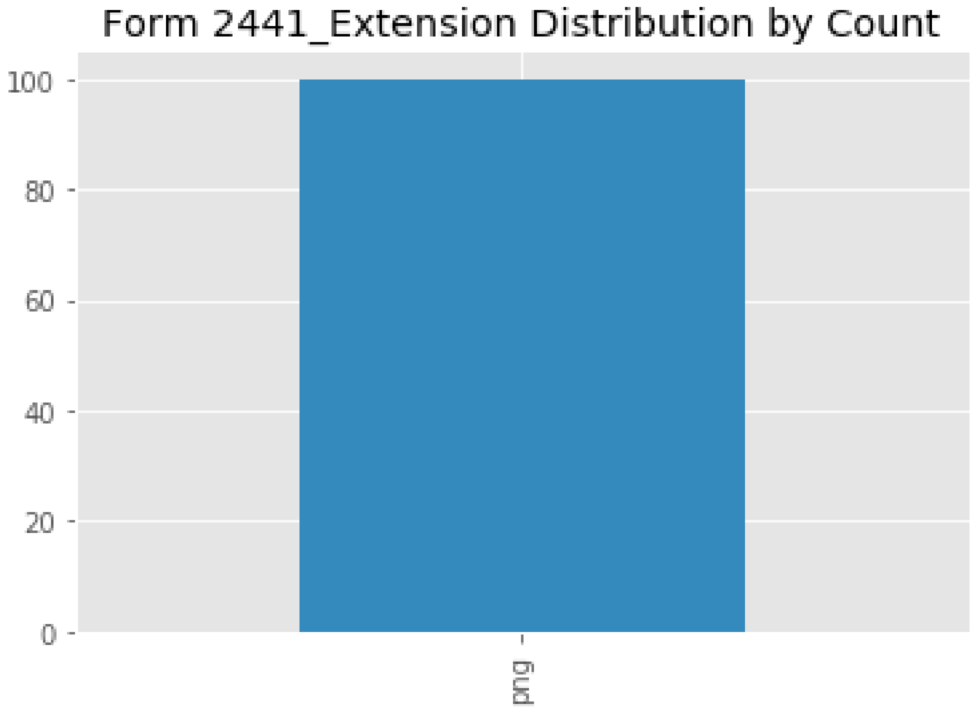


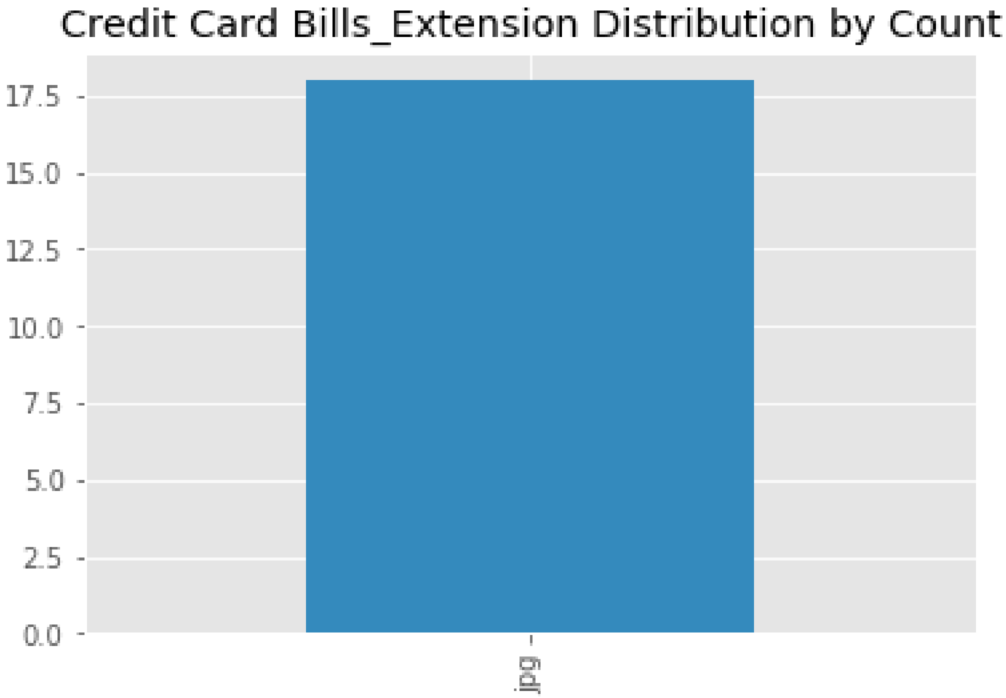

Insurance_Extension Distribution by Count
30
25
20
15
10
5
0
tif
jpg

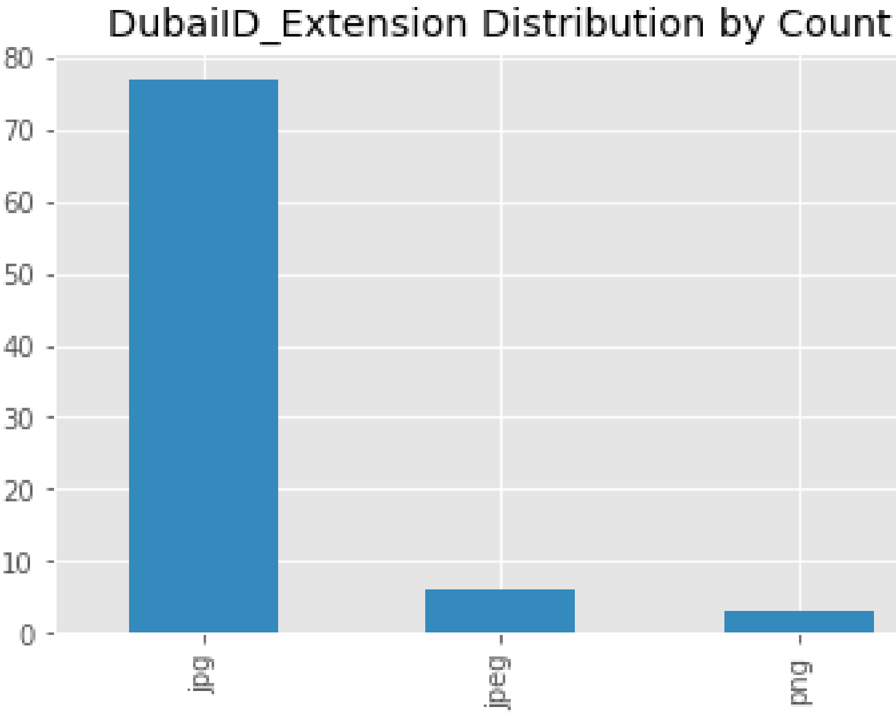

Resume_Extension Distribution by Count
120
100
80
60
40
20
0
jpg

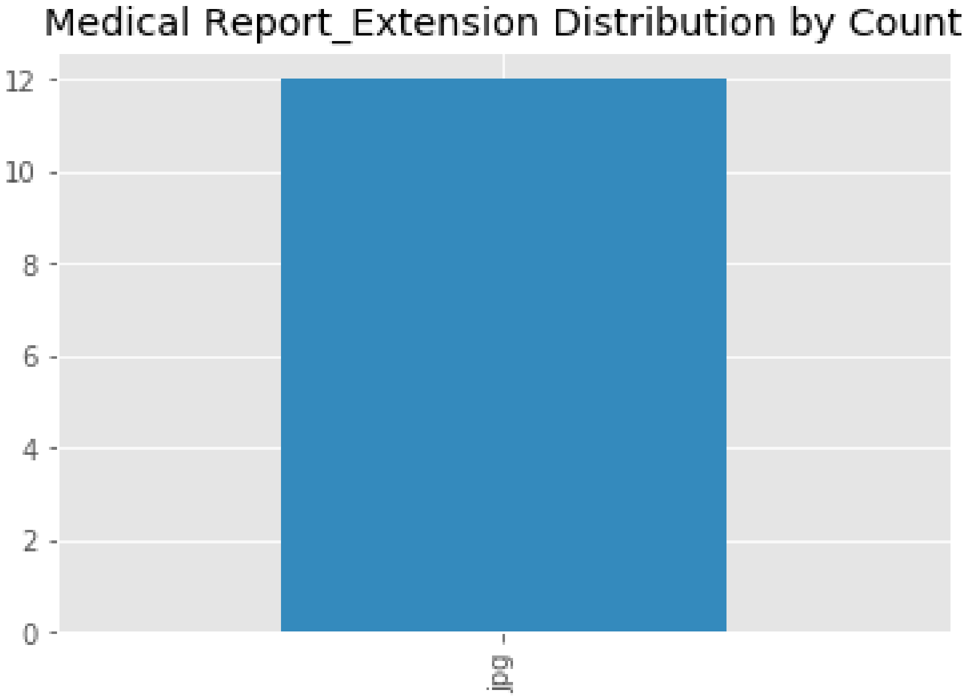

Aadhar_Extension Distribution by Count
80
60
40
20
0
jpg
jpeg
JPG

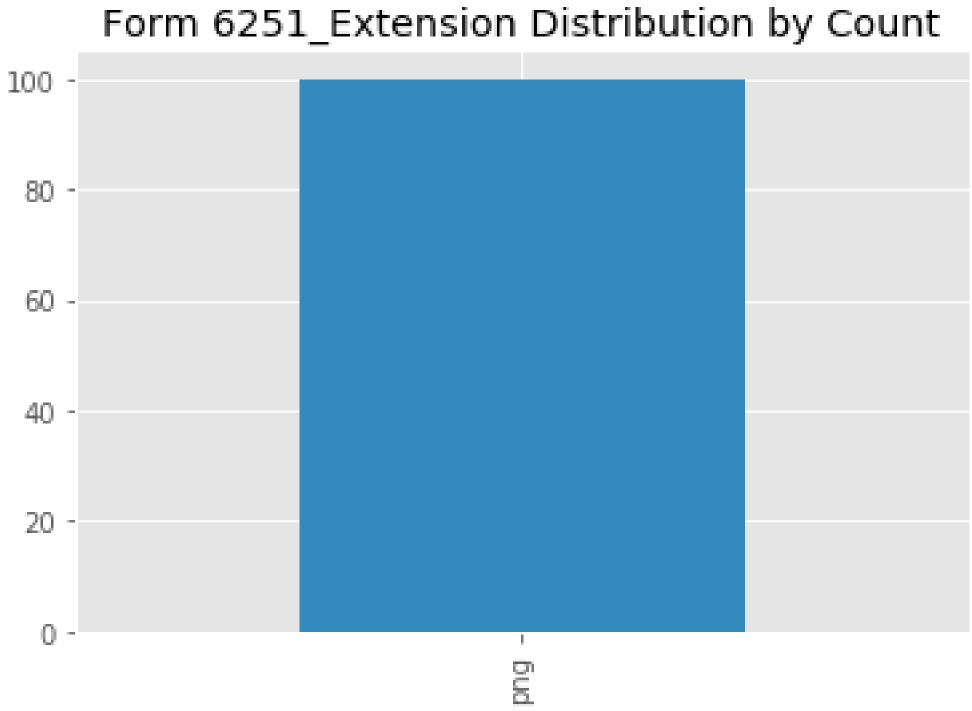

Form 1040_Extension Distribution by Count
100
80
60
40
20
0
png

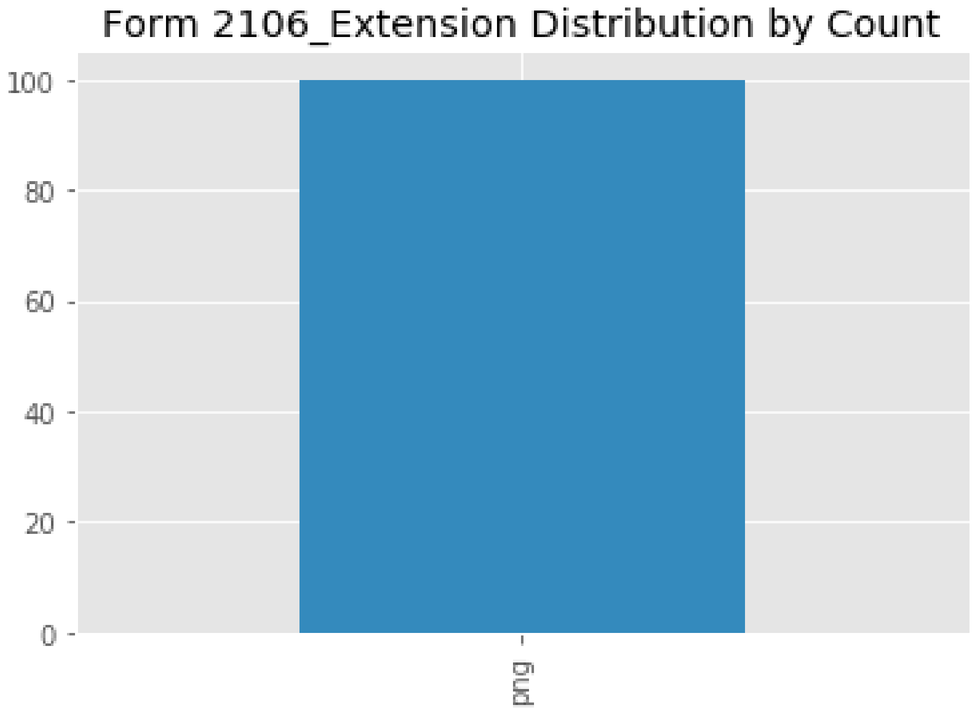

Form 2441_Extension Distribution by Count
100
80
60
40
20
0
png

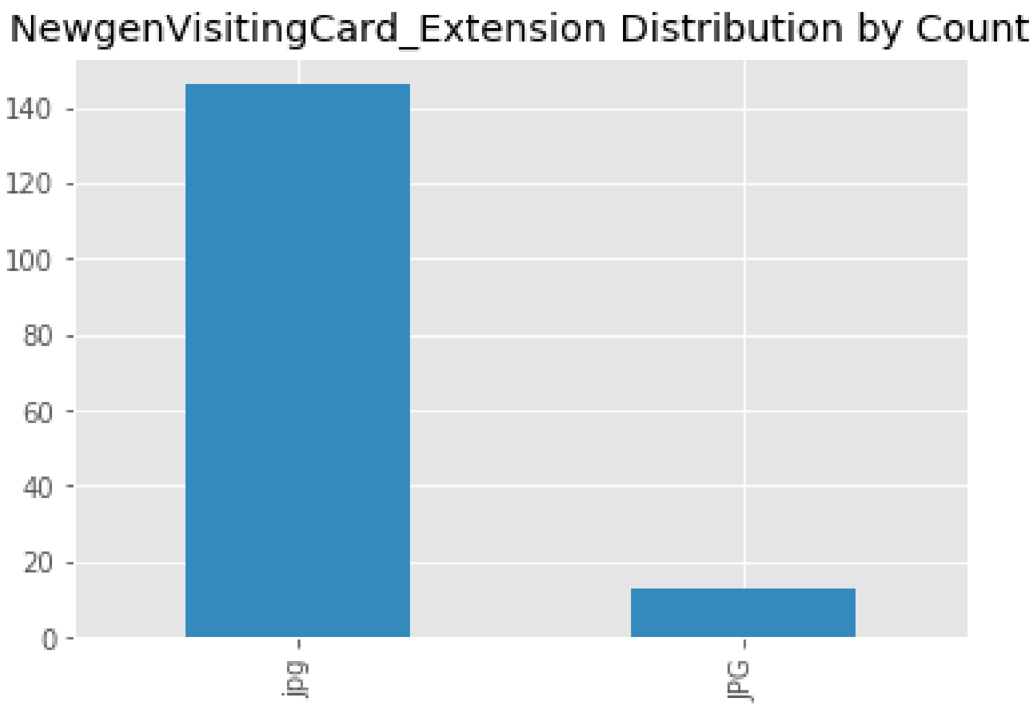

Credit Card Bills_Extension Distribution by Count
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
jpg

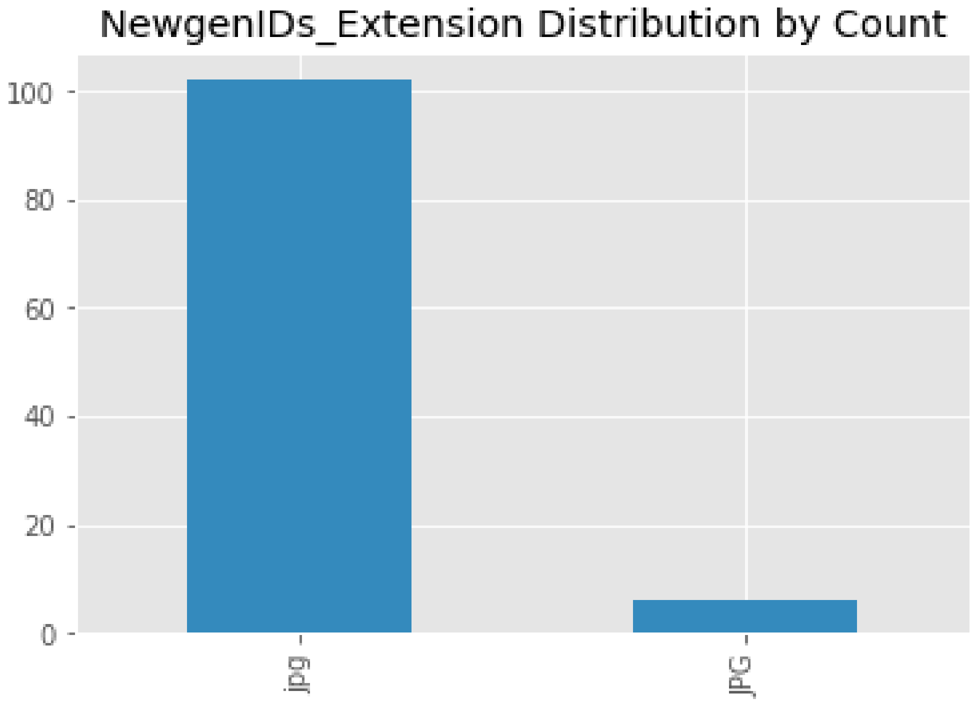

DubaiID_Extension Distribution by Count
80
70
60
50
40
30
20
10
0
jpg
jpeg
png

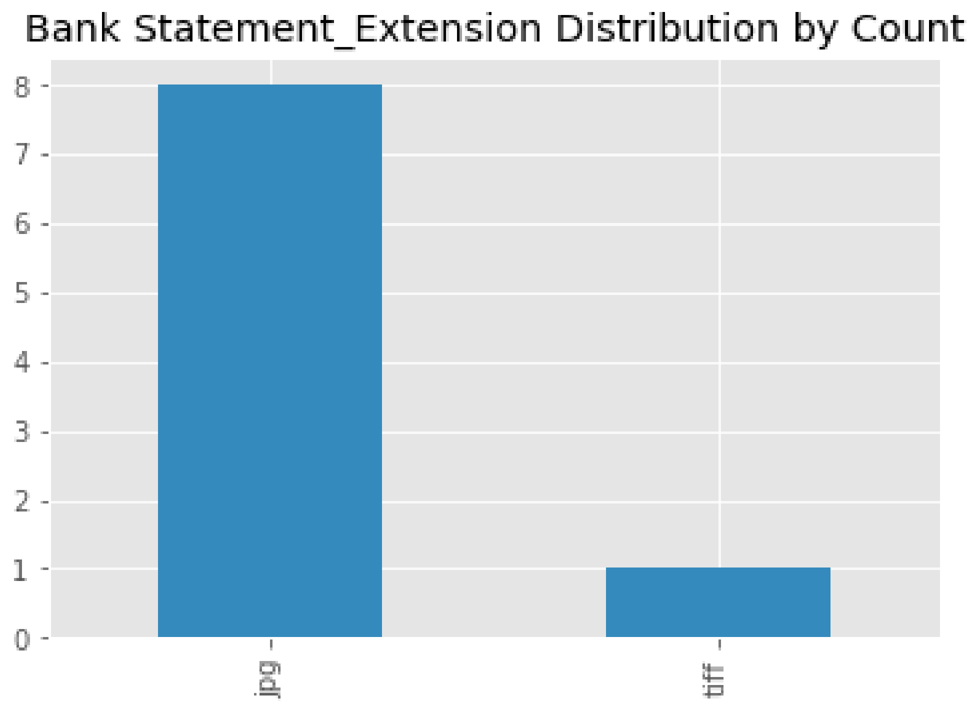

Medical Report_Extension Distribution by Count
12
10
8
6
4
2
0
jpg

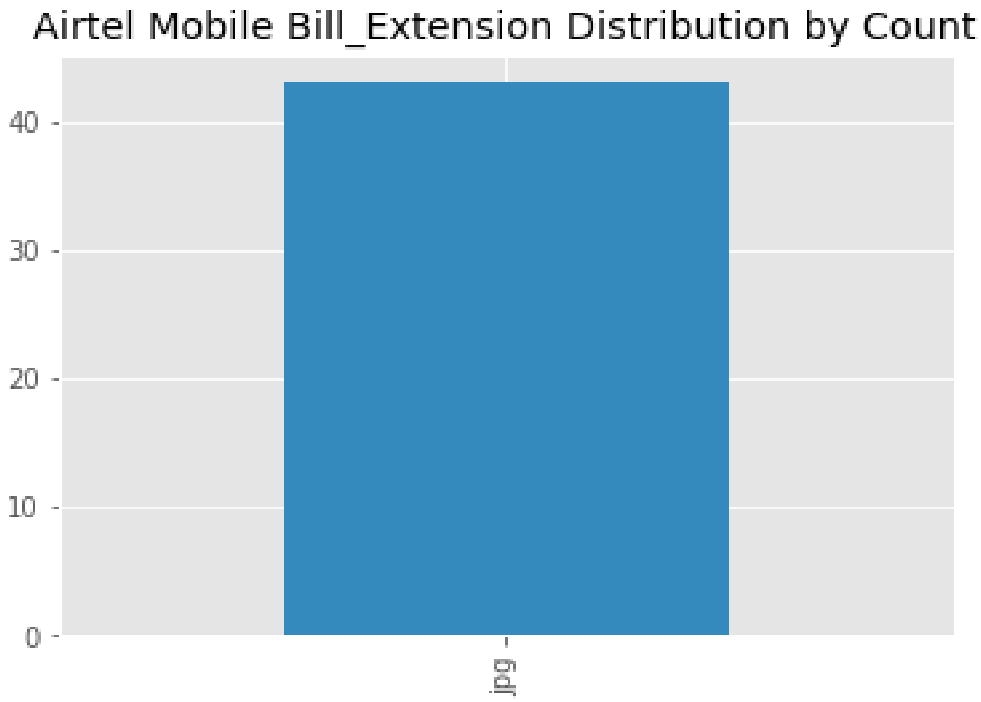

Form 6251_Extension Distribution by Count
100
80
60
40
20
0
png

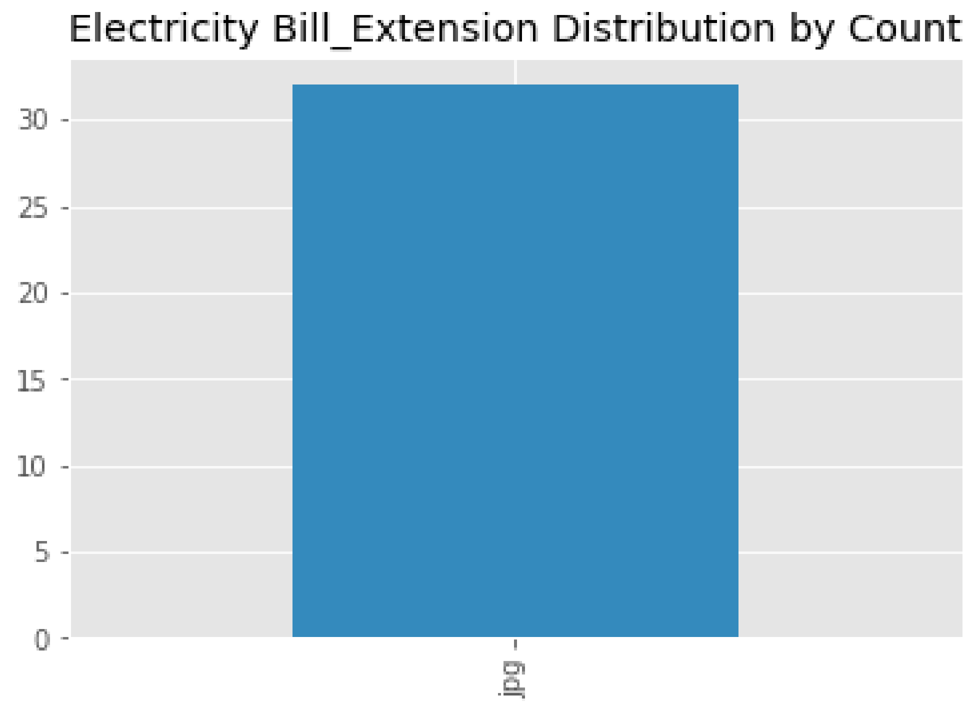

Form 2106_Extension Distribution by Count
100
80
60
40
20
0
png

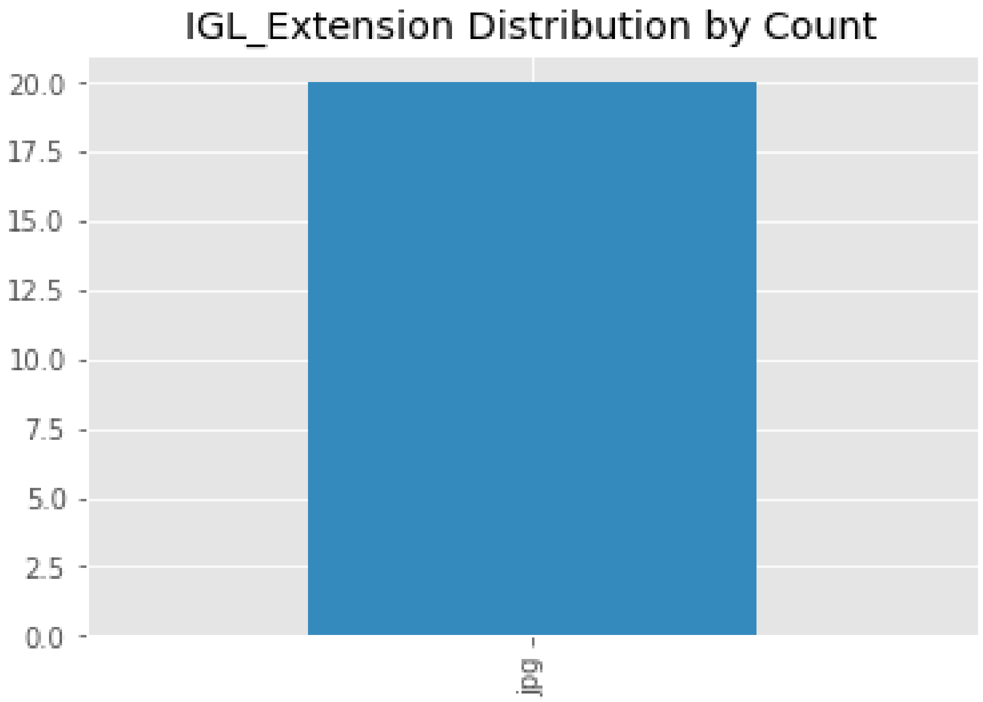

NewgenVisitingCard_Extension Distribution by Count
140
120
100
80
60
40
20
0
jpg
JPG

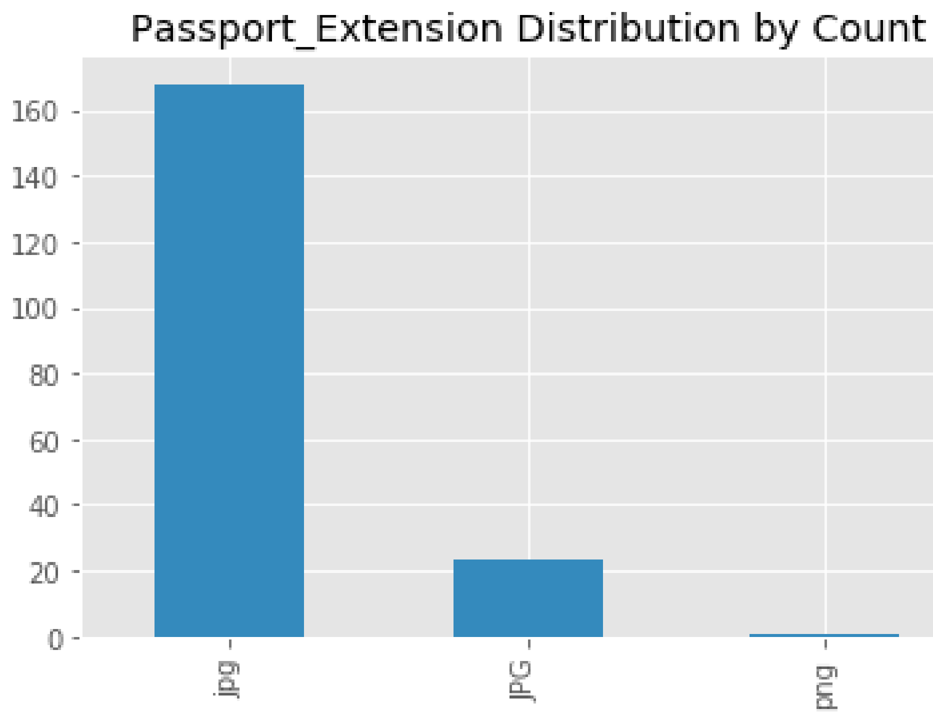

NewgenIDs_Extension Distribution by Count
100
80
60
40
20
0
jpg
JPG


Bank Statement_Extension Distribution by Count
0
1
2
3
4
5
6
7
8
jpg
tiff


Airtel Mobile Bill_Extension Distribution by Count
0
10
20
30
40
jpg


Electricity Bill_Extension Distribution by Count
30
25
20
15
10
5
0
jpg

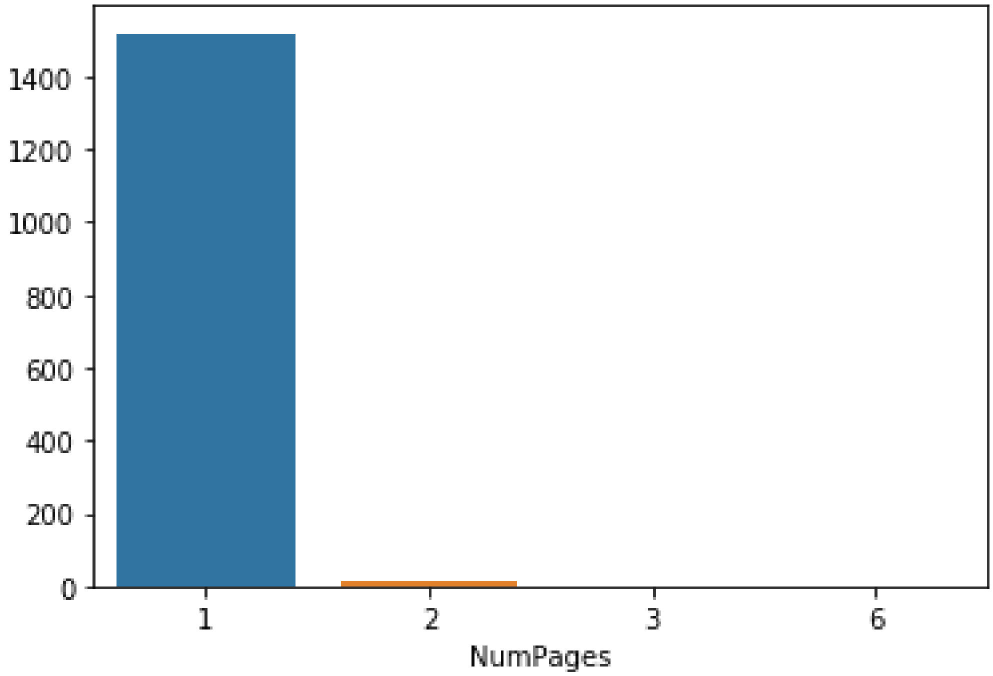

IGL_Extension Distribution by Count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
jpg



```
<Figure size 432x288 with 0 Axes>
```

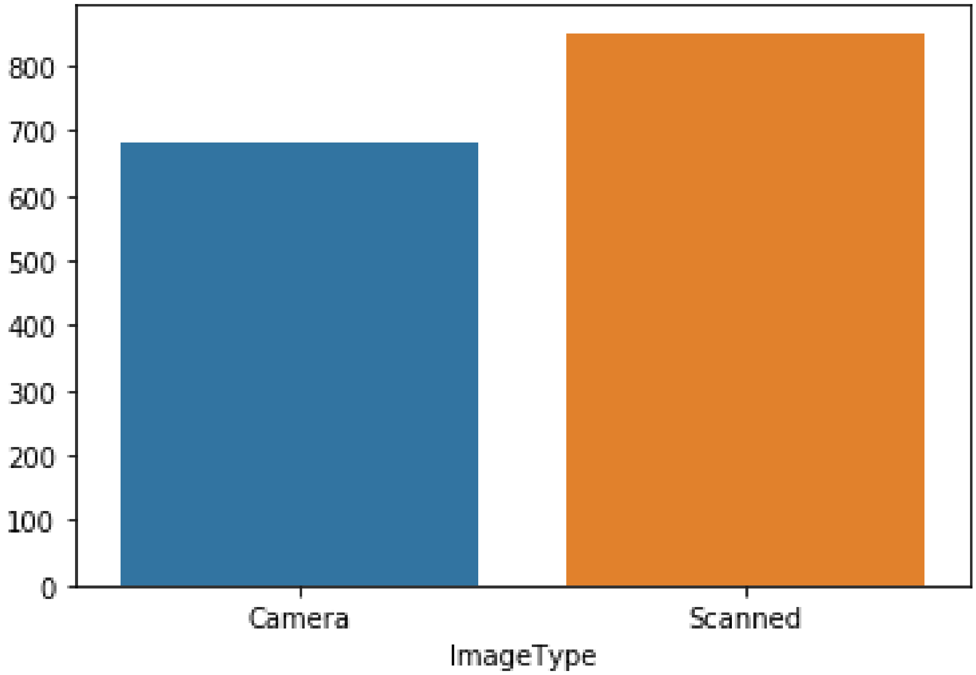
7 COUNT NUMBER OF PAGES IN EACH FILE

```
[21]: <matplotlib.axes._subplots.AxesSubplot at 0x7f66f75c2a90>
```



8 Checking Camera Capured or Scanned

```
[26]: <matplotlib.axes._subplots.AxesSubplot at 0x7f66f75c2490>
```



9 Finding Image Quality

Image Quality of each file has been saved in the **pre_data_report.csv** file

```
[27]:                                              FileName ClassName  NumPages  \
      0  /home/abhinav/dataset_analysis/image_clusterin…       PAN         1
      1  /home/abhinav/dataset_analysis/image_clusterin…       PAN         1
      2  /home/abhinav/dataset_analysis/image_clusterin…       PAN         1
      3  /home/abhinav/dataset_analysis/image_clusterin…       PAN         1
      4  /home/abhinav/dataset_analysis/image_clusterin…       PAN         1

         ImageType      FileSize  ColorDepth    imageDpi
      0     Camera  (3456, 4608)           8    (72, 72)
      1     Camera  (4608, 3456)           8    (72, 72)
      2    Scanned    (983, 611)           8  (300, 300)
      3    Scanned             0           0           0
      4     Camera  (4032, 3024)           8    (72, 72)
```