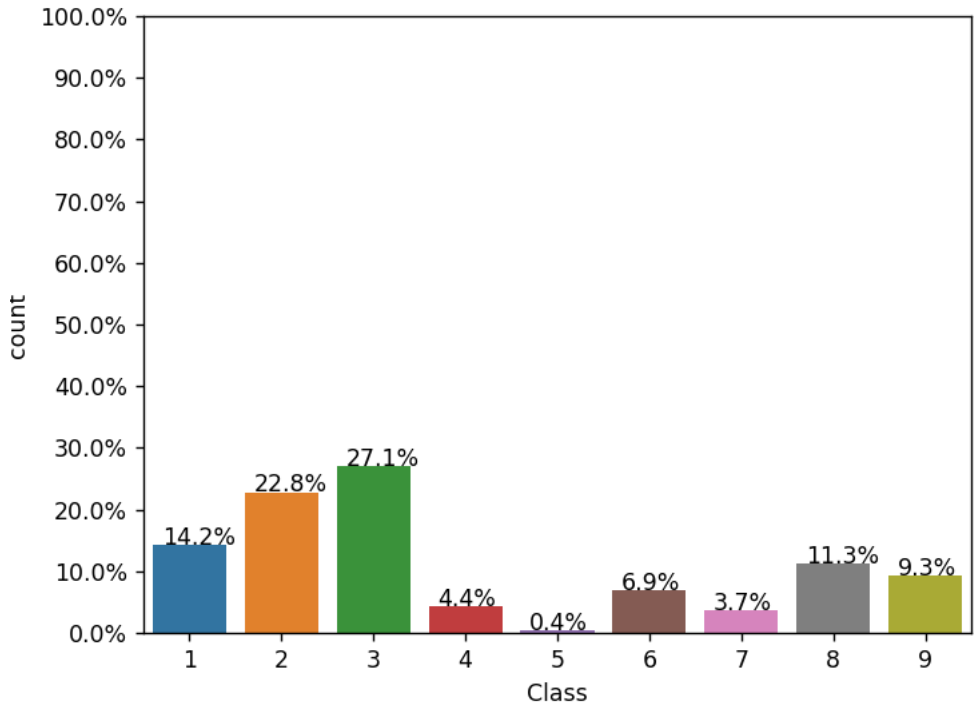
Exploratory Data Analysis

Created	@Jun 10, 2021 12:31 PM
Tags	

Bytes file

Number of data points in each class

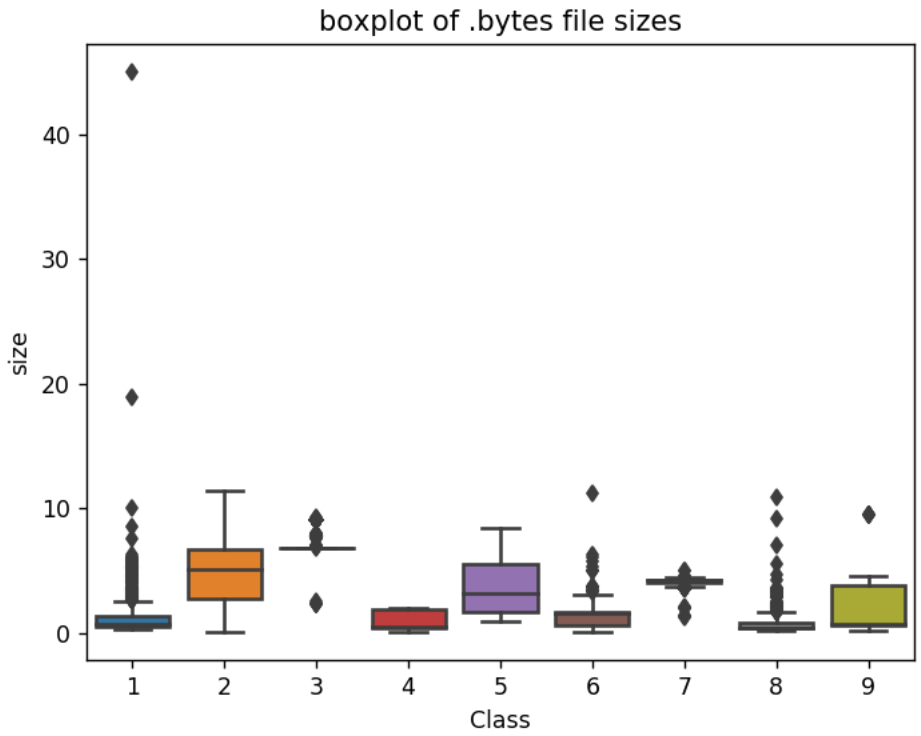


Observation:- Class 5 has less number of data points, **Imbalance data problem.**

File size as feature

ID	File Name	Size
0	01azqd4InC7m9JpocGv5	4.234863
1	01IsoiSMh5gxyDYTI4CB	5.538818
2	01jsnpXSAIgw6aPeDxrU	3.887939
3	01kcPWA9K2BOxQeS5Rju	0.574219
4	01SuzwMJEIXsK7A8dQbl	0.370850

Box plot of file size as feature



Observation:- Class 2, 5 and 9 can be easily distinguished from other classes, using only the file size feature

Copy of Bag of word as feature of the file

# ID	Aa File Name	# 0	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	...	# f9	# fa	# fb	# fc
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758
1	01IsoiSMh5gxyDYTI4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133
4	01SuzwMJEIXsK7A8dQbl	19764	710	302	433	559	410	262	249	422	...	350	209	239	653

Copy of Combining Bag of Words and File size as Features

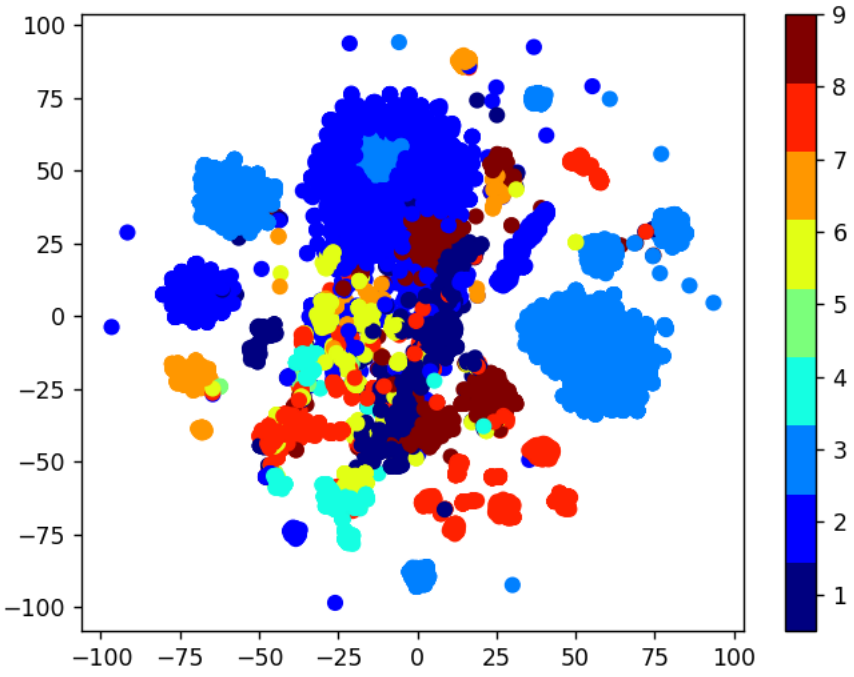
# ID	Aa File Name	# 0	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	...	# f9	# fa	# fb	# fc
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758
1	01IsoiSMh5gxyDYTI4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133
4	01SuzwMJEIXsK7A8dQbl	19764	710	302	433	559	410	262	249	422	...	350	209	239	653

Copy of Normalizing the Features

ID	File Name	0	1	2	3	4	5	6	7	8
0	01azqd4InC7m9JpocGv5	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.002946	0.002638
1	01IsoiSMh5gxyDYTI4CB	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.006984	0.008267
2	01jsnpXSAlgw6aPeDxrU	0.040827	0.013434	0.001429	0.001315	0.005464	0.00528	0.005078	0.002155	0.008104
3	01kcPWA9K2BOxQeS5Rju	0.009209	0.001708	0.000404	0.000441	0.00077	0.000354	0.00031	0.000481	0.000959
4	01SuzwMJElXsK7A8dQbl	0.008629	0.001	0.000168	0.000234	0.000342	0.000232	0.000148	0.000229	0.000376

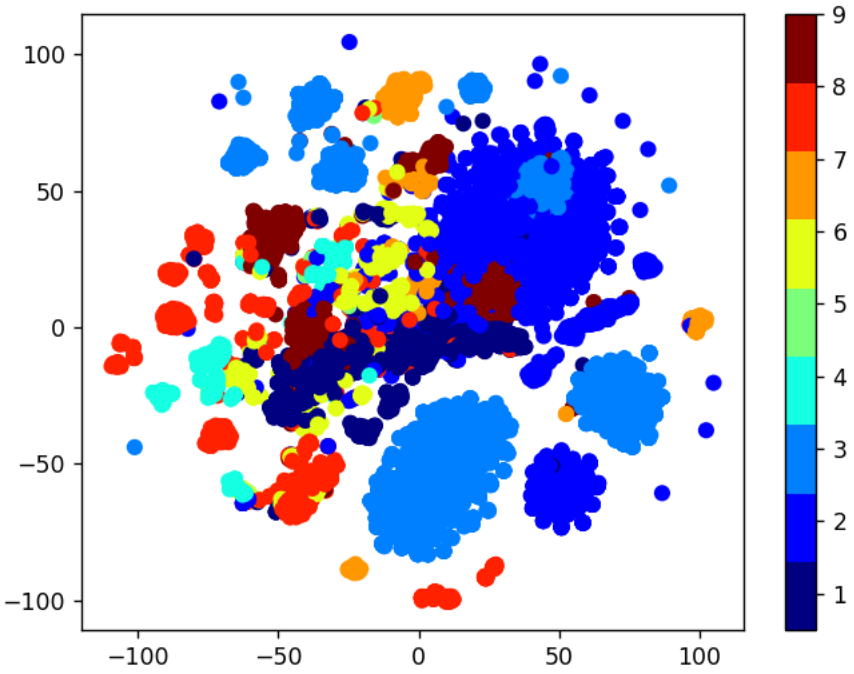
Multivariate analysis of the Features

Perplexity = 50



Observation:- Class 2 and 3 are clearly separated whereas other classes have poor distinctions

Perplexity = 30



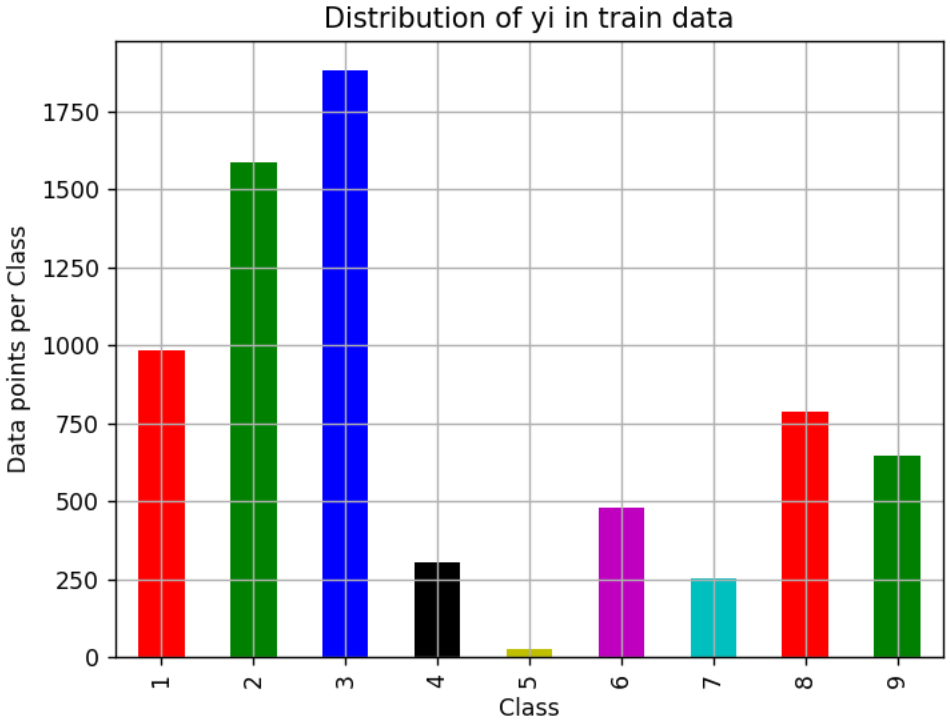
Observation:- Class 2 and 3 are clearly separated whereas other classes have poor distinctions

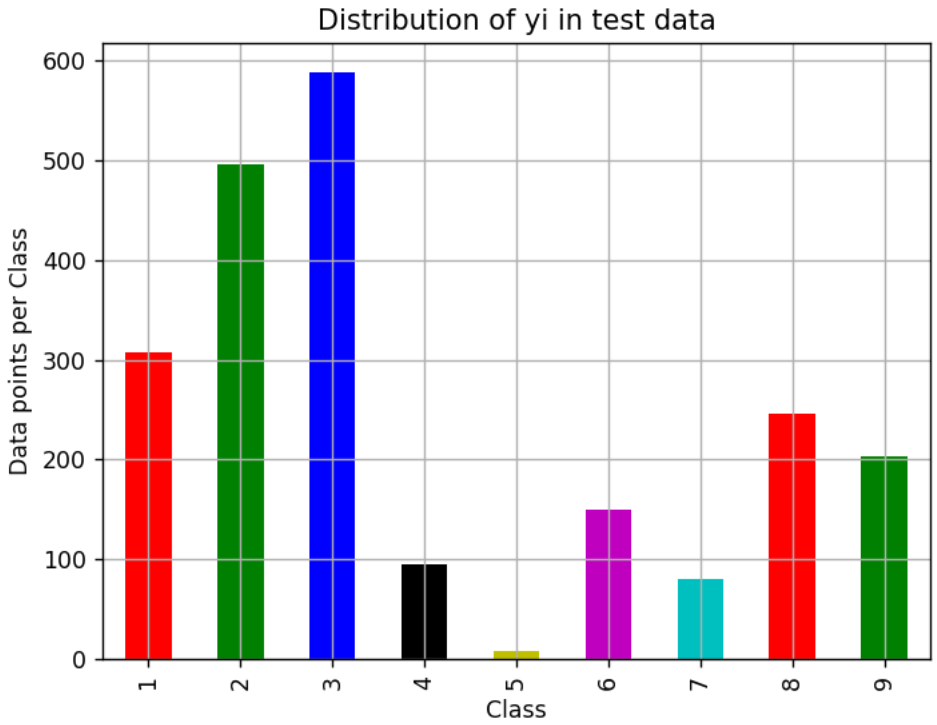
Test Train Split

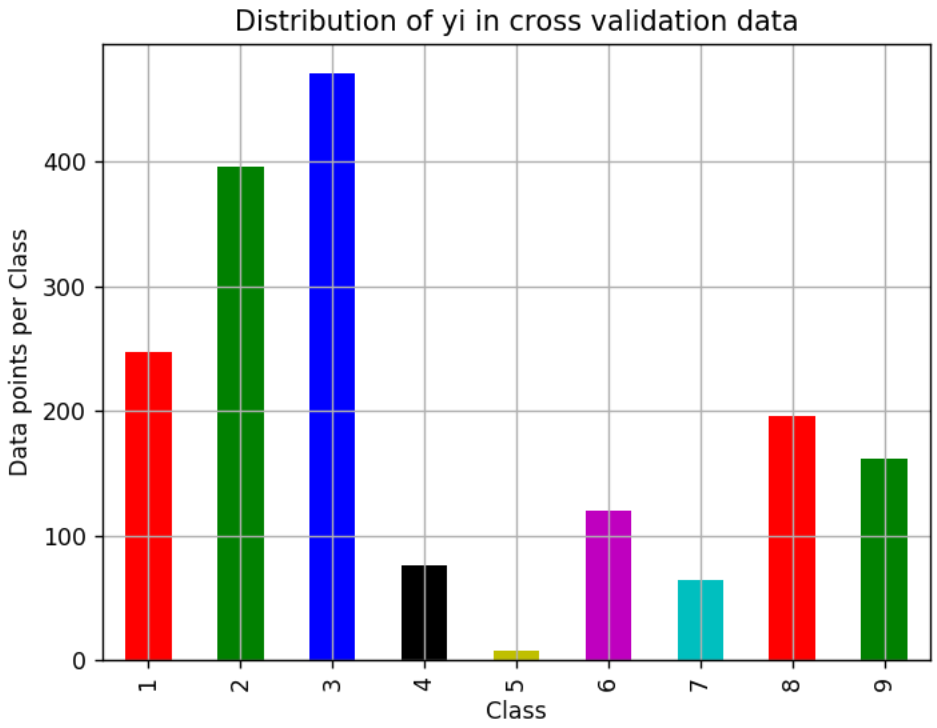
Number of data points in train data: 6955
Number of data points in test data: 2174
Number of data points in cross validation data: 1739

Check for distribution of data

We check for the distribution of classes in each split by plotting a histogram.


Distribution of yi in train data
Data points per Class
0
250
500
750
1000
1250
1500
1750
1
2
3
4
5
6
7
8
9
Class


Distribution of yi in test data
Data points per Class
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9
Class

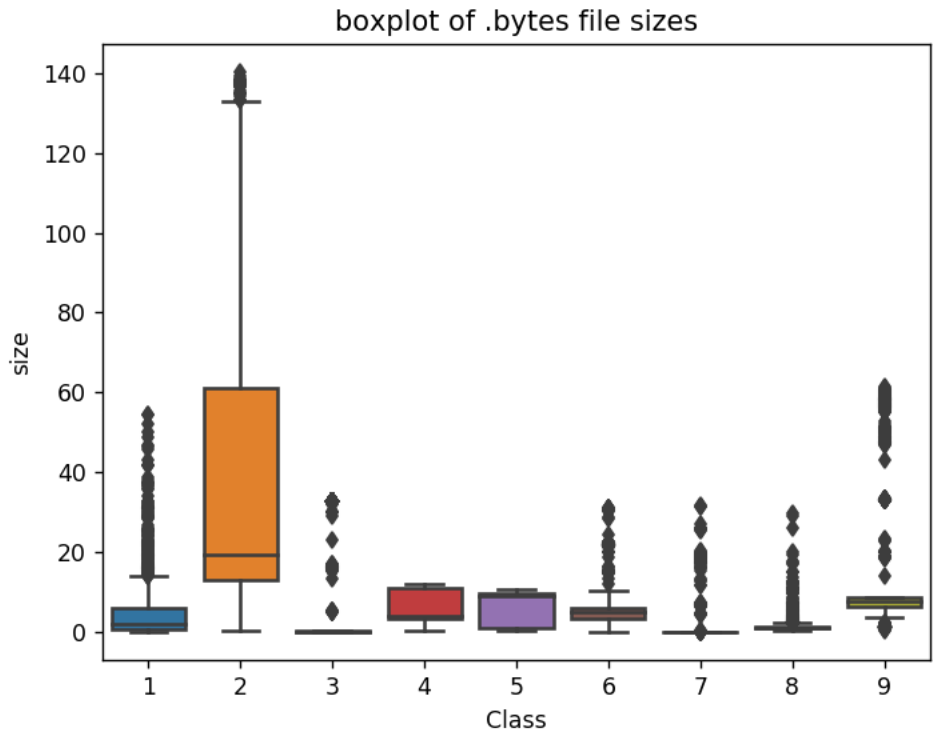


ASM file

File Size as feature

ID	File Name	Size
0	01azqd4InC7m9JpocGv5	56.229886
1	01IsoiSMh5gxyDYTI4CB	13.999378
2	01jsnpXSAlgw6aPeDxrU	8.507785
3	01kcPWA9K2BOxQeS5Rju	0.078190
4	01SuzwMJEIXsK7A8dQbl	0.996723

Box plot of file size as feature



Copy of Bag of Words for ASM File

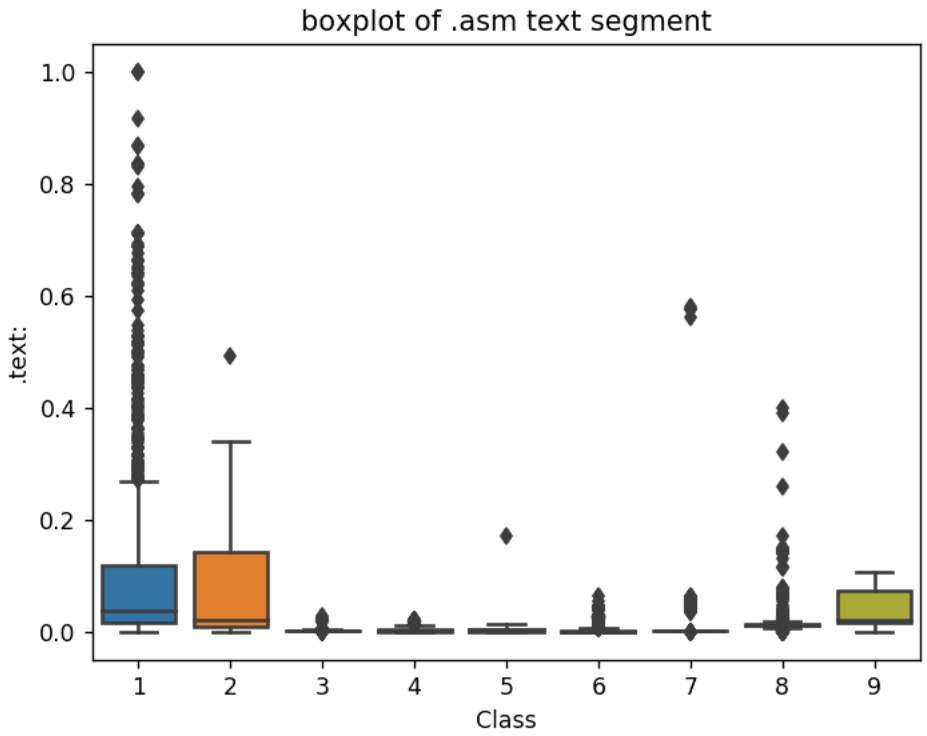
Property	File Name	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	edx	esi	eax
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	18	66	15
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	18	29	48
2	3ekVow2ajZHbTnBcsDfX	17	427	0	50	43	0	145	0	3	...	13	42	10
3	3X2nY7iQaPBlWDrAZqJe	17	227	0	43	19	0	0	0	3	...	6	8	14
4	46OZzdsSKDCFV8h7XWxf	17	402	0	59	170	0	0	0	3	...	12	9	18

Copy of Combining Bag of Words and File size as Features

ID	File Name	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	esi	eax	ebx	ecx
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	66	15	43	83
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	29	48	82	12
2	3ekVow2ajZHbTnBcsDfX	17	427	0	50	43	0	145	0	3	...	42	10	67	14
3	3X2nY7iQaPBlWDrAZqJe	17	227	0	43	19	0	0	0	3	...	8	14	7	2
4	46OZzdsSKDCFV8h7XWxf	17	402	0	59	170	0	0	0	3	...	9	18	29	5

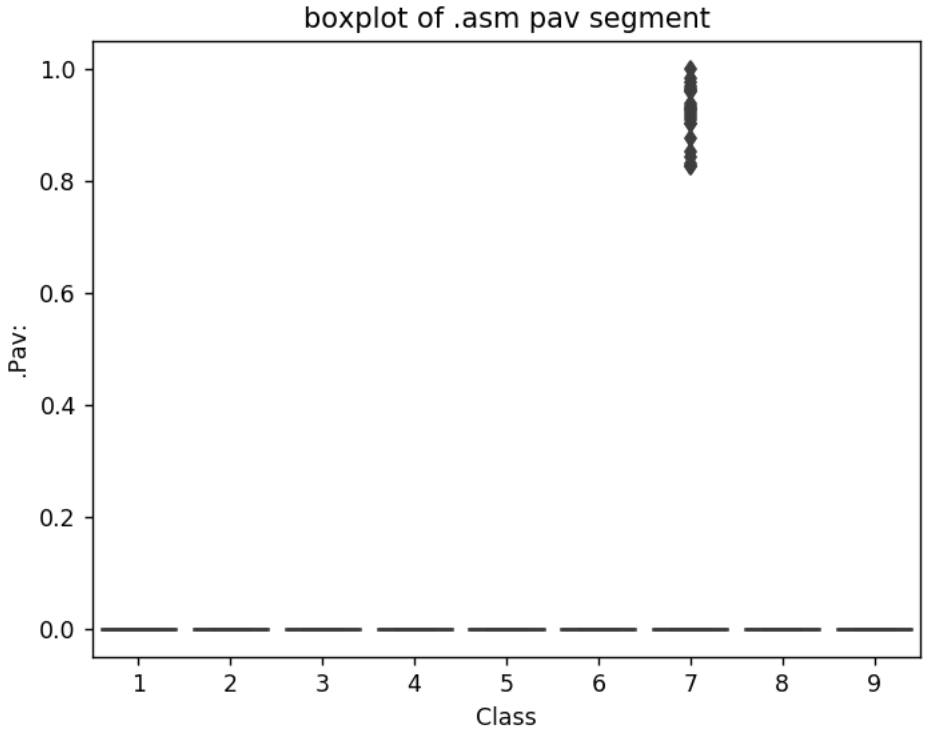
Univariate Analysis

.text vs Class



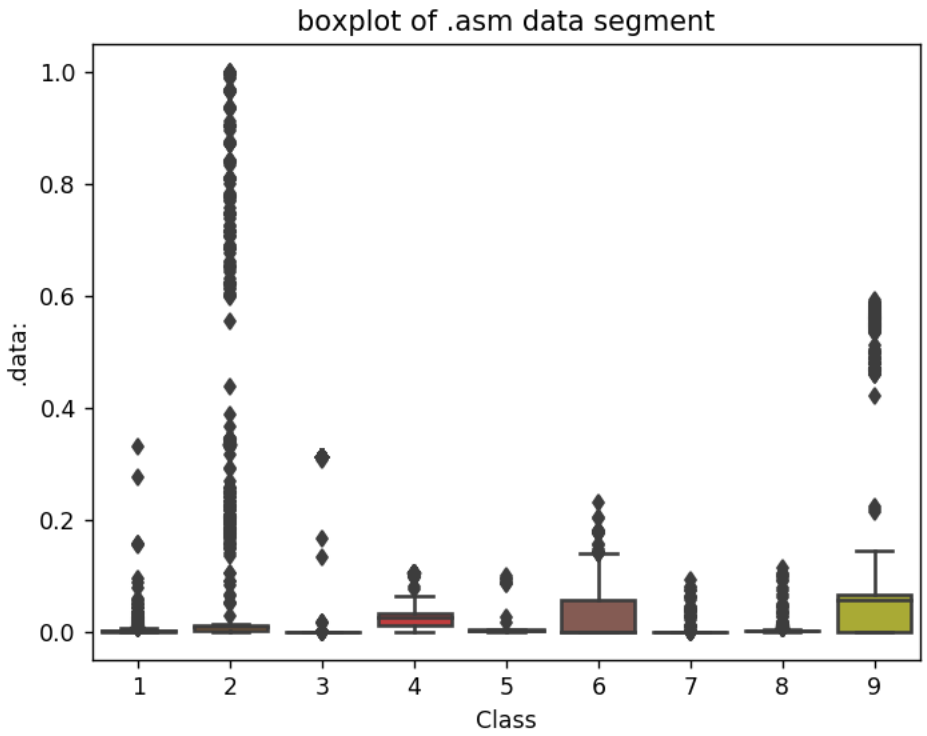

```
The plot is between Text and class
Class 1,2 and 9 can be easly separated
```

.Pav vs Class



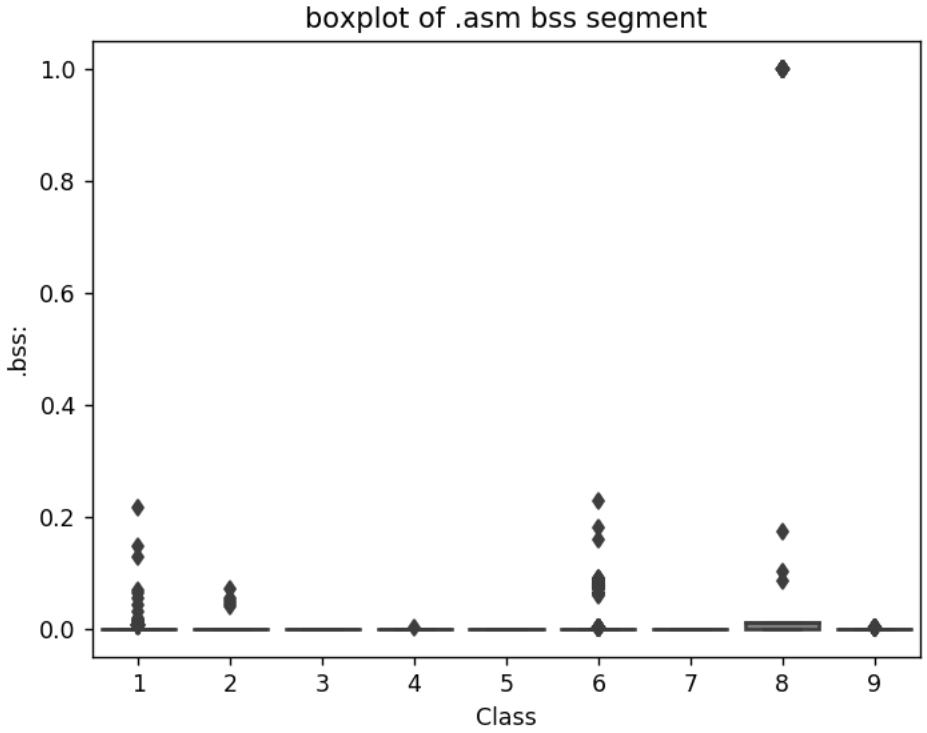

```
No useful information
```

.data vs Class



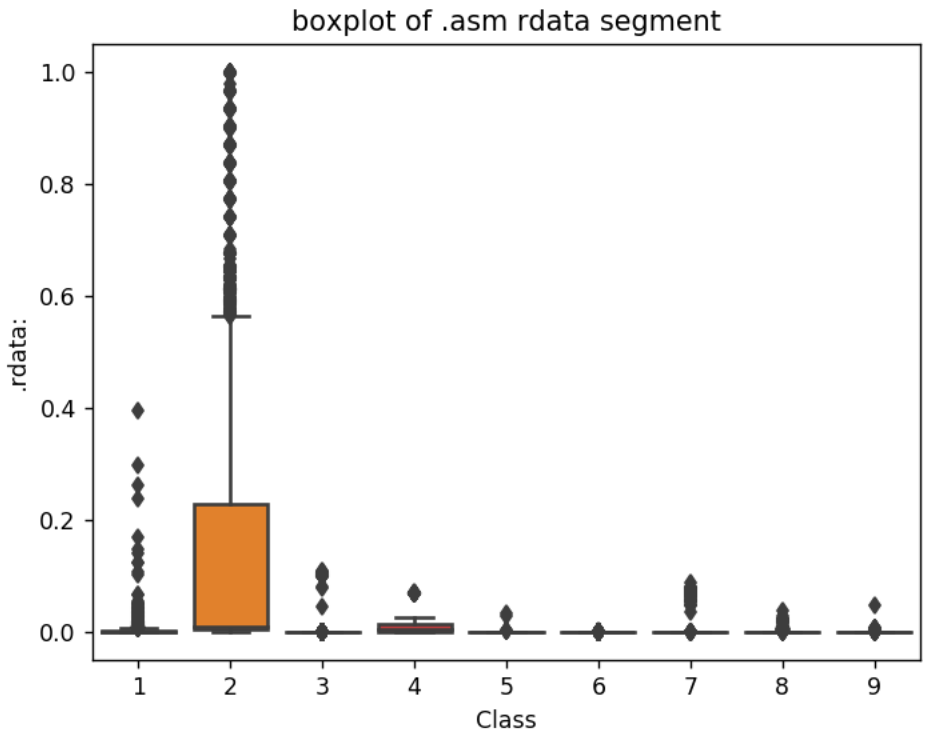

```
The plot is between data segment and class label
class 6 and class 9 can be easily separated from given points
```

.bss vs Class



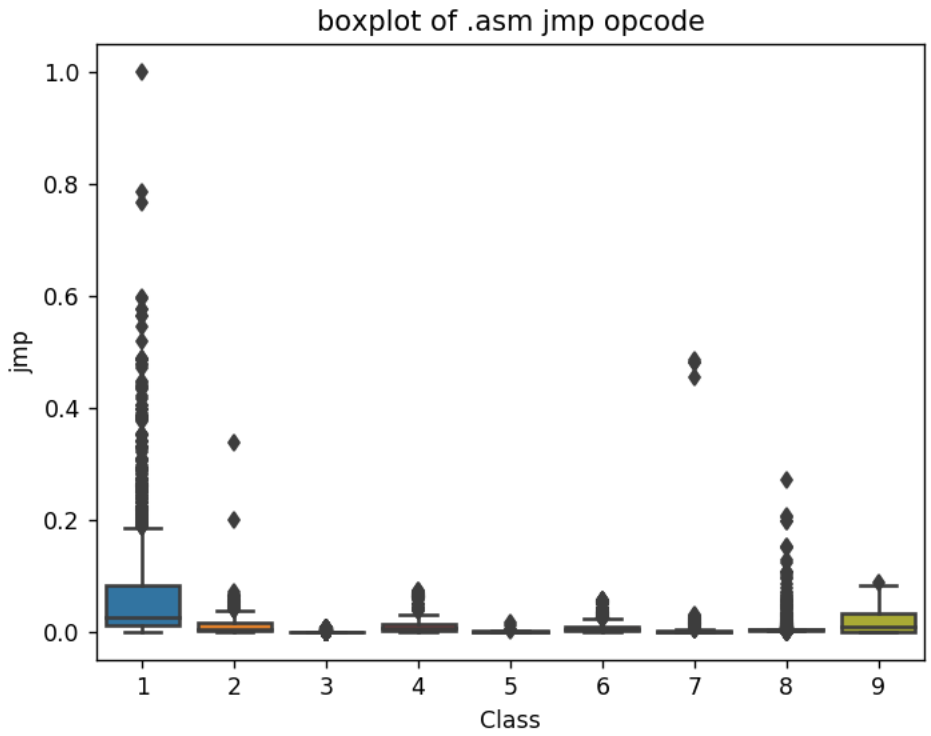

```
plot between bss segment and class label
very less number of files are having bss segment
```

.rdata vs Class



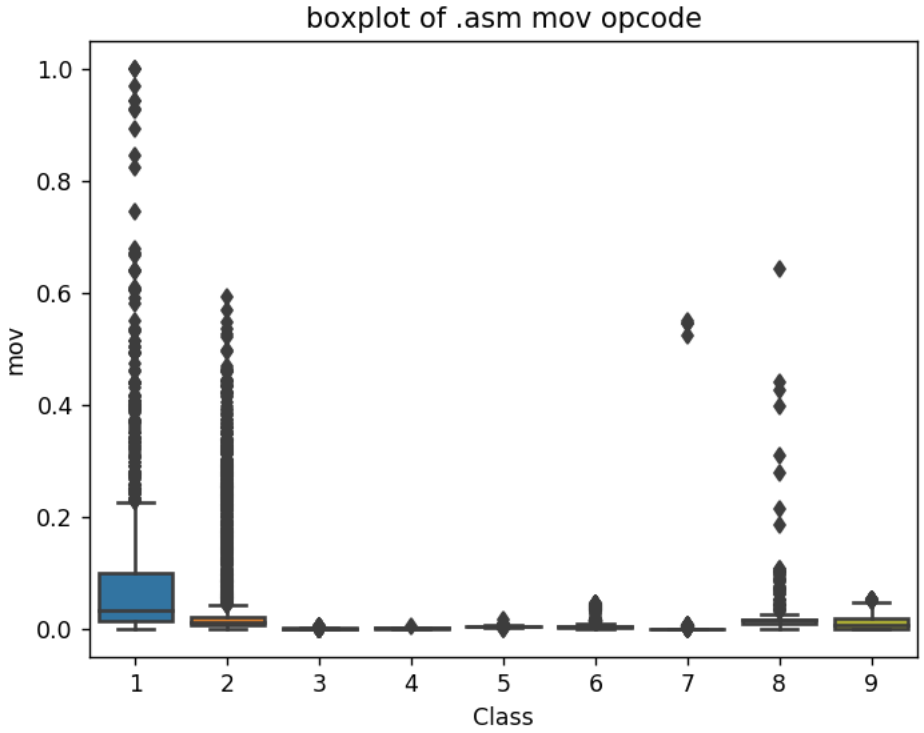

```
Plot between rdata segment and Class segment
Class 2 can be easily separated 75 pecentile files are having 1M rdata lines
```

jmp vs Class



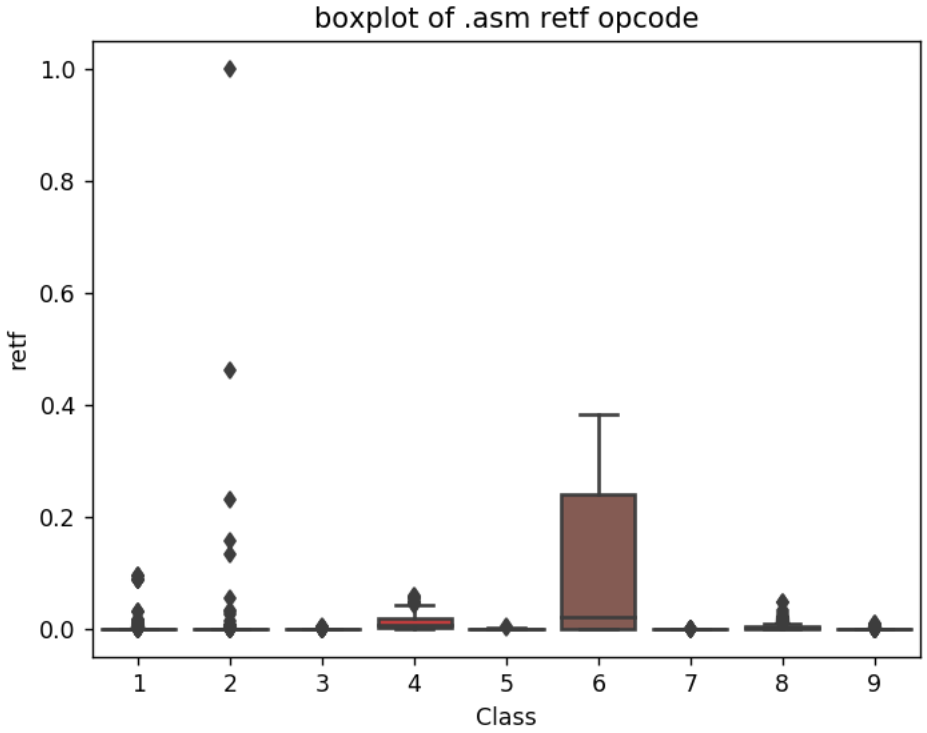

```
plot between jmp and Class label
Class 1 is having frequency of 2000 approx in 75 perentile of files
```

mov vs Class



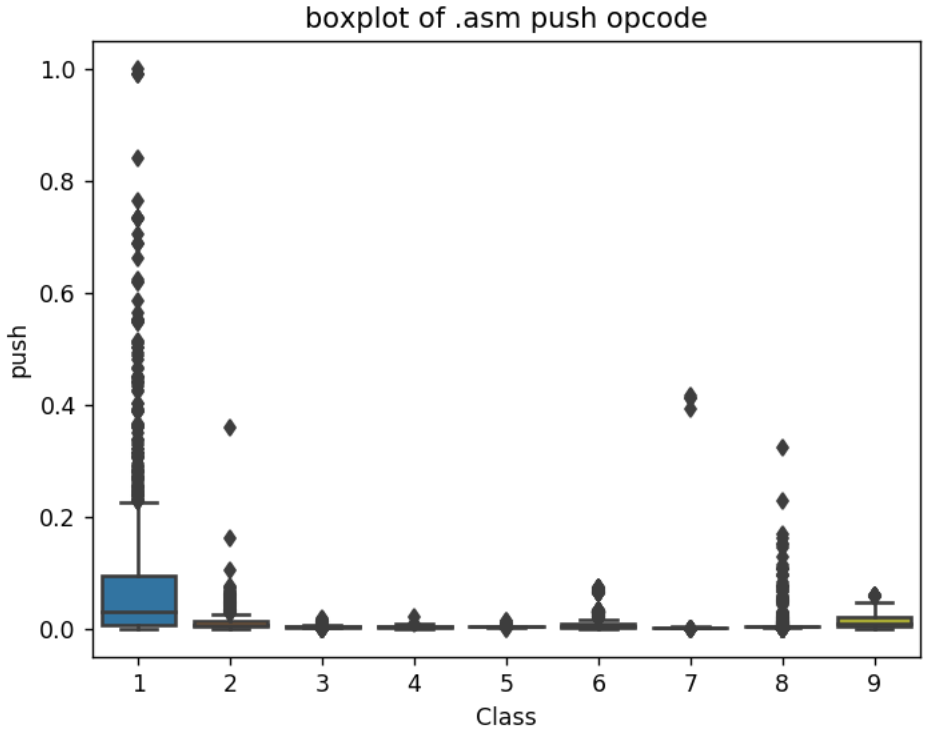

```
plot between Class label and mov opcode
Class 1 is having frequency of 2000 approx in 75 perentile of files
```

retf vs Class




```
plot between Class label and retf
Class 6 can be easily separated with opcode retf
The frequency of retf is approx of 250.
```

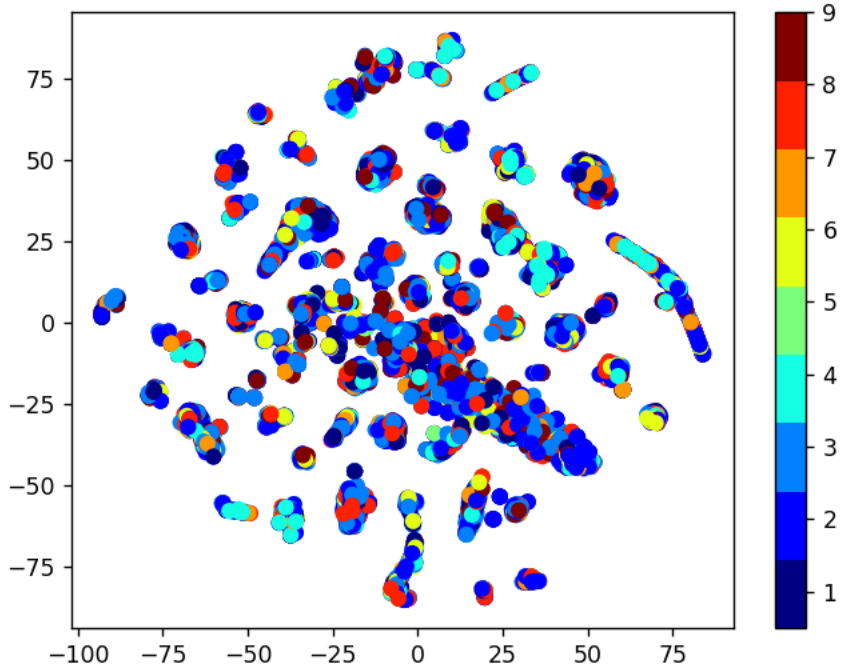
push vs Class



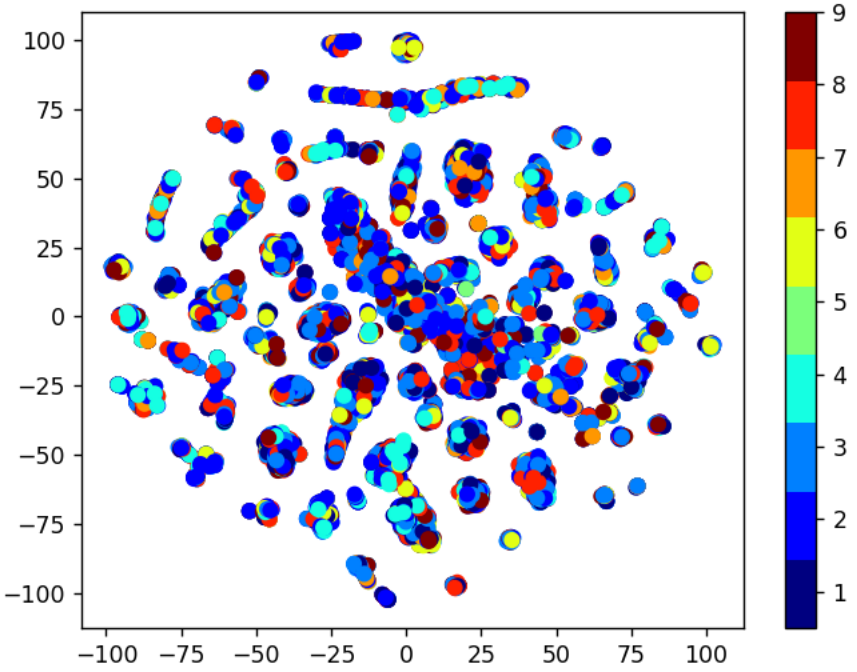

```
plot between push opcode and Class label
Class 1 is having 75 precentile files with push opcodes of frequency 1000
```

Multivariate Analysis

Perplexity = 50



Perplexity = 30



Conclusion of EDA

- We have taken only 52 features from asm files (after reading through many blogs and research papers)
- The univariate analysis was done only on few important features.
- Take-aways
  1. Class 3 can be easily separated because of the frequency of segments, opcodes and keywords being less
  2. Each feature has its unique importance in separating the Class labels.

Train and test Split

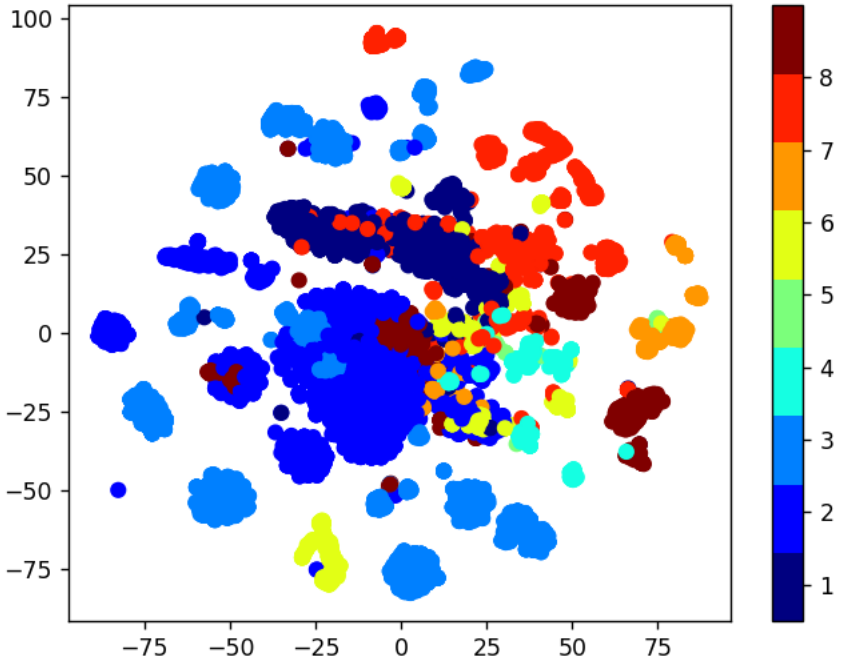
Number of data points in train data: 6955
Number of data points in test data: 2174
Number of data points in cross validation data: 1739

Merged features

Merging both asm and byte file features

ID	0	1	2	3	4	5	6	7	8	9	...	edx
0	0.262806	0.005498	0.001567	0.002067	0.002048	0.001835	0.002058	0.002946	0.002638	0.003531	...	0.0154
1	0.017358	0.011737	0.004033	0.003876	0.005303	0.003873	0.004747	0.006984	0.008267	0.000394	...	0.0049
2	0.040827	0.013434	0.001429	0.001315	0.005464	0.00528	0.005078	0.002155	0.008104	0.002707	...	0.0000
3	0.009209	0.001708	0.000404	0.000441	0.00077	0.000354	0.00031	0.000481	0.000959	0.000521	...	0.0003
4	0.008629	0.001	0.000168	0.000234	0.000342	0.000232	0.000148	0.000229	0.000376	0.000246	...	0.0003

Multivariate Analysis on final features



Train and test Split

Number of data points in train data: 6955
Number of data points in test data: 2174
Number of data points in cross validation data: 1739