Letter Recognition EDA

Matt Oehler

March 12, 2018

1 Description

Being able to autonomously read handwritten documents would be instrumental in saving time for several businesses, and revolutionizing the process of things such as family history. However, since this is a non-trivial task due to the wide variety of handwriting, statisical methods can be used to train a computer to read documents instead of having to do it manually.

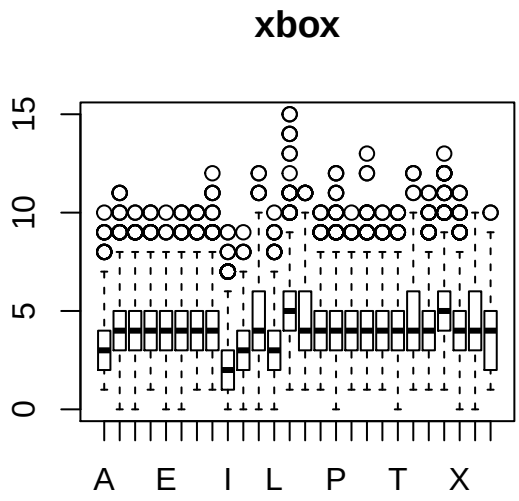
2 Data

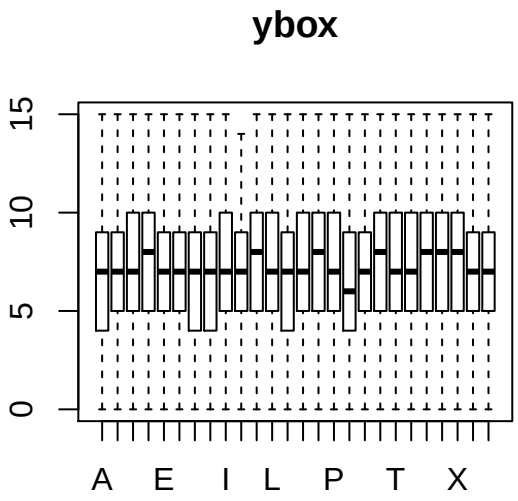
For this problem we have 16 attirbutes of letters contained in a historical document. The figures below show the structure of the data set, the frequency of each letter occuring in the dataset, and the side by side box plots for each of the covariates for each letter. Some variables vary quite a bit between the letters, but some variables such as 'high' have little variance between the letters

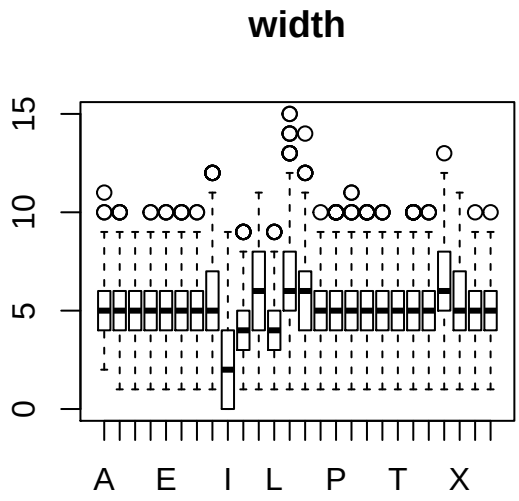
	letter	xbox	ybox	width	high	pix	xbar	ybar
1	I	5	12	3	7	2	10	5
2	D	4	11	6	8	6	10	6
3	N	7	11	6	6	3	5	9
4	G	2	1	3	1	1	8	6
5	S	4	11	5	8	3	8	8
6	B	4	2	5	4	4	8	7

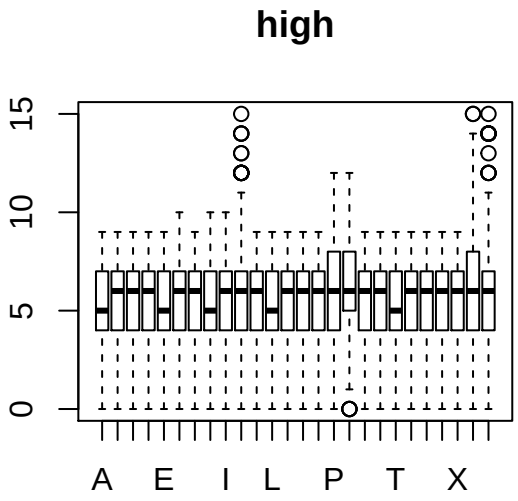
	x2bar	y2bar	xybar	x2ybar	xy2bar	xege	xegevy	yege
1	5	4	13	3	9	2	8	4
2	2	6	10	3	7	3	7	3
3	4	6	4	4	10	6	10	2
4	6	6	6	5	9	1	7	5
5	6	9	5	6	6	0	8	9
6	6	6	7	6	6	2	8	7

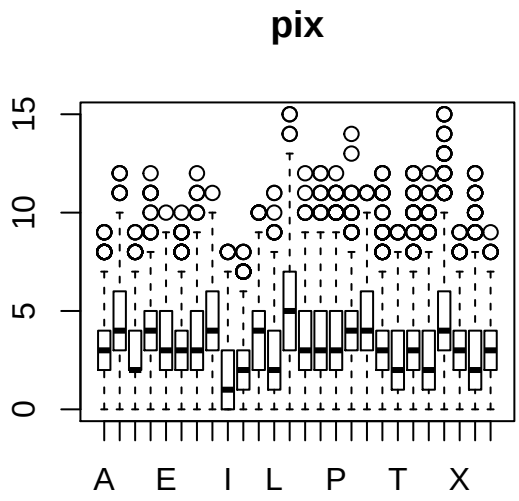
```
  A   B   C   D   E   F   G   H   I   J   K   L   M   N   O   P   Q   R
789 766 736 805 768 775 773 734 755 747 739 761 792 783 753 803 783 758
  S   T   U   V   W   X   Y   Z
748 795 813 764 752 787 786 734
```

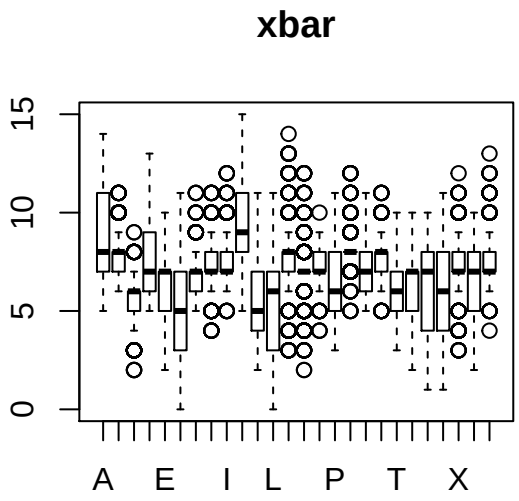

xbox
0
5
10
15
A
E
I
L
P
T
X

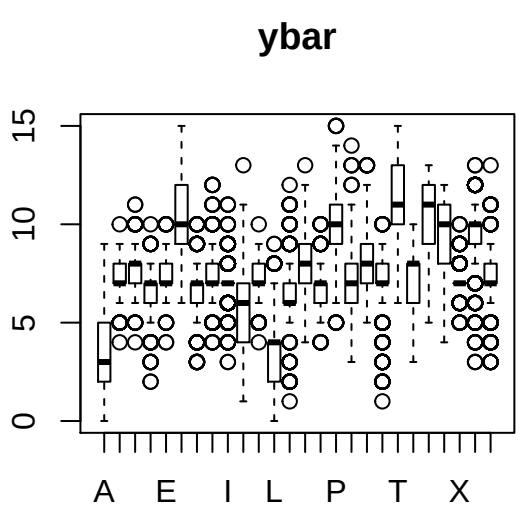

ybox
0
5
10
15
A
E
I
L
P
T
X

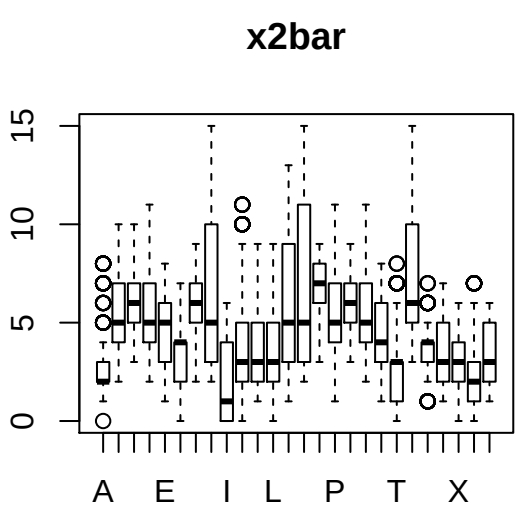

width
0
5
10
15
A
E
I
L
P
T
X

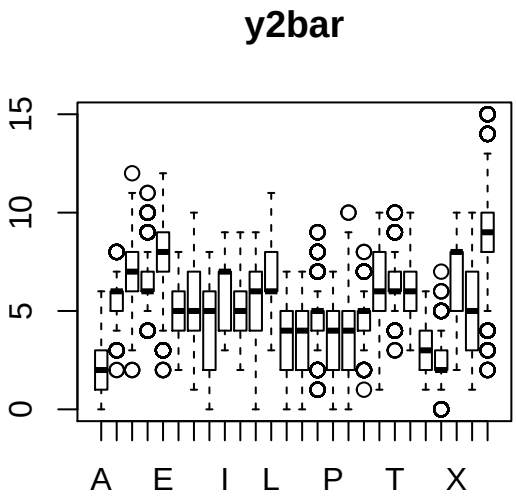

high
0
5
10
15
A
E
I
L
P
T
X

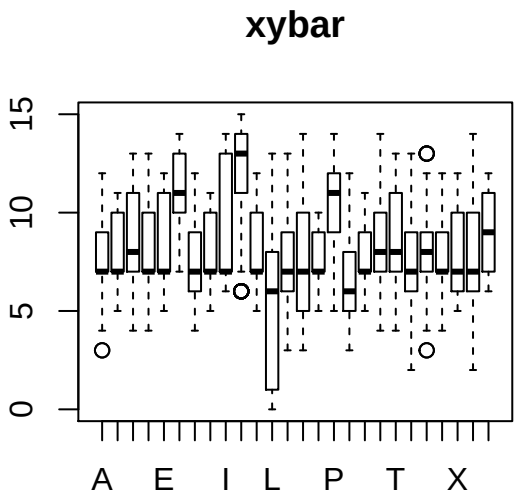

pix
0
5
10
15
A
E
I
L
P
T
X

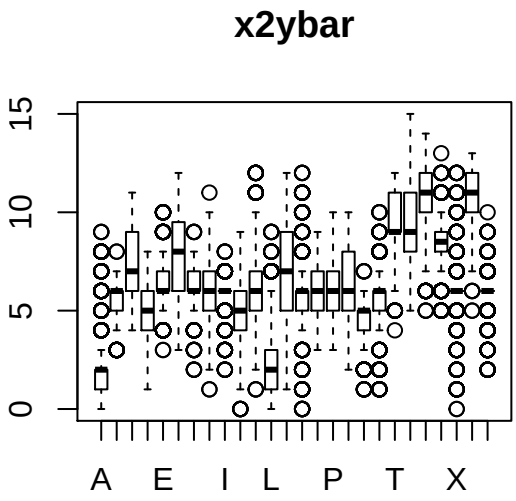

xbar
0
5
10
15
A
E
I
L
P
T
X

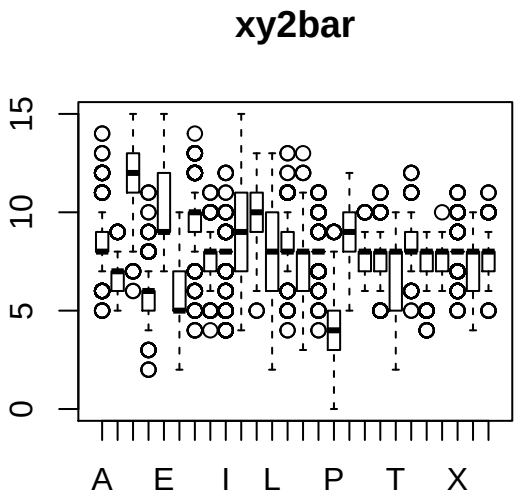

ybar
0
5
10
15
A
E
I
L
P
T
X

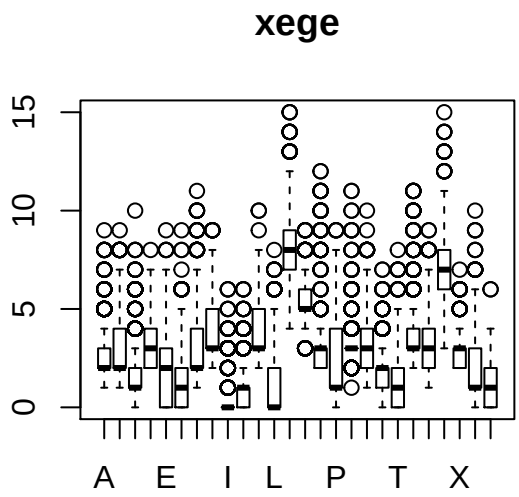

x2bar
0
5
10
15
A
E
I
L
P
T
X

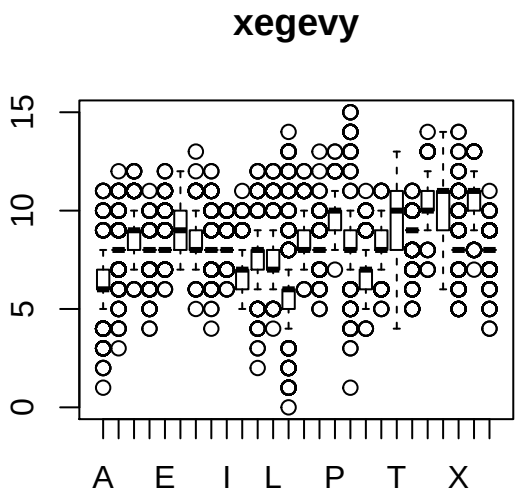

y2bar
0
5
10
15
A
E
I
L
P
T
X

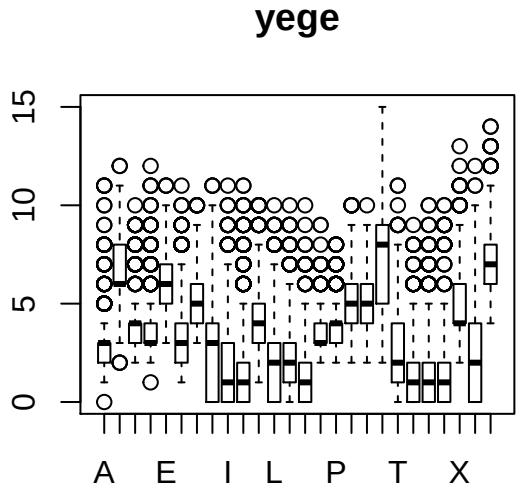

xybar
0
5
10
15
A
E
I
L
P
T
X

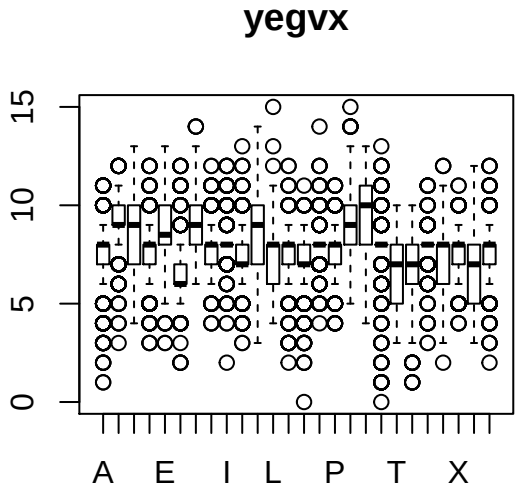

x2ybar
0
5
10
15
A
E
I
L
P
T
X


xy2bar
0
5
10
15
A
E
I
L
P
T
X









3 Method

I think that a method such as support vector machines or neural networks would be appropriate for this problem since they can handle classification problems when the response has more than two classes.

4 Something I don't know

I don't know of other methods to approach multi-level classification problems. I am also unsure what would happen if we had data where there weren't a balanced level of classes.