Khan Inan
Mid-term project

2a.
Yes the plot of the average values does still represent the digit label, just in less detail. This is because when R is plotting the values normally, the averages are essentially already shown in density of pixels in a given area. The plot of the average, simply takes these more concentrated areas of pixels and puts a single dot in those areas, where there would otherwise be many.

2.b
The digits that fare better under this operation are the more simplet digits such as 1 and 7, and this is because the average of the values sort of scramble together with the more winding and complex digits like 6,8,9.

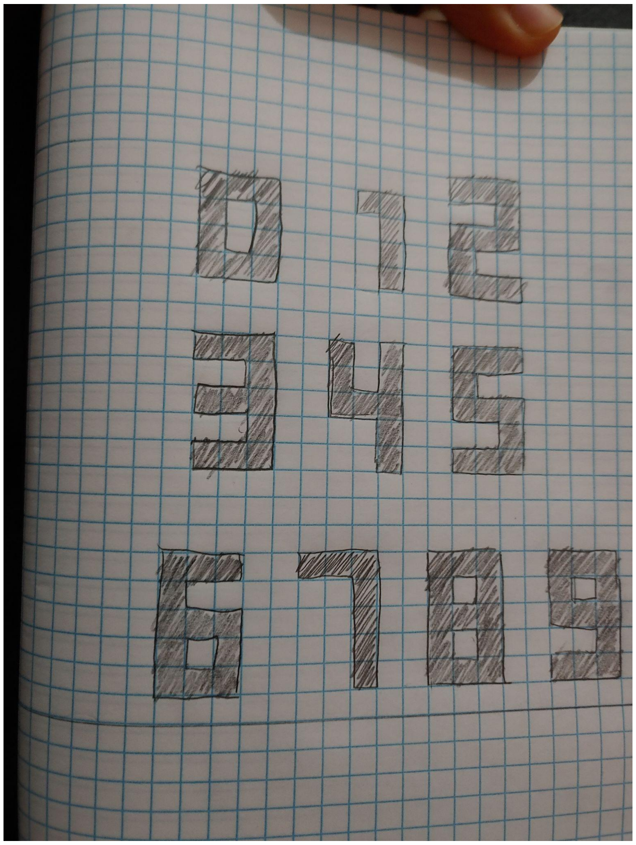
3.a
The columns that have the highest variance are

3.b
Yes I can connect the variances to the results in 2b and this is because columns that go through more digits have more variance due to having to fill in more parts of each digit. The columns that do not go through any digits or very few have either 0 or very little variance

3.c
Yes, changing the columns with the lowest variability average will have an impact on the digits because to make each digit more pronounced, you need many dots that can create the shape of the digit. Lowering the overall # of these values will make the digits less pronounced and easily visible.

3.d
There are no columns that have an average value of close to 255 and this is because there would need to be a solid line going through the entire plot in order for the average of the values in any column to come to 255. 0 on the other hand is much more common and this is due to all the blank spaces between digits on the graph, or columns where values are sparsely populated

4.

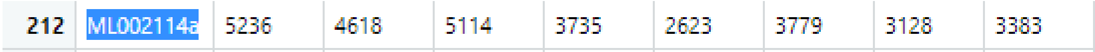


1

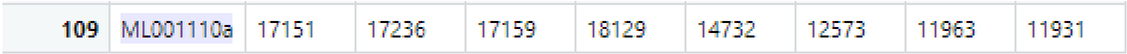
The top 5 genes with the highest average expression are ML000314a, ML00062a, ML000719a, ML001110a and ML002114a

2

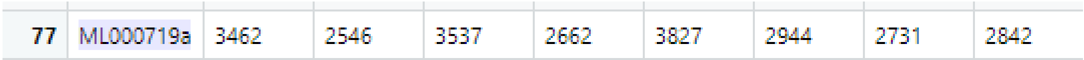
No, the top 5 genes do not seem to be different because it seems as though expression is consistently high for all the categories for genes that have high average expression



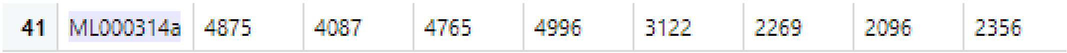
212	ML002114a	5236	4618	5114	3735	2623	3779	3128	3383



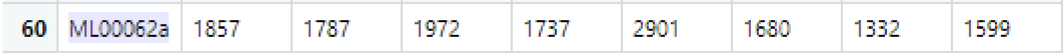
109	ML001110a	17151	17236	17159	18129	14732	12573	11963	11931



77	ML000719a	3462	2546	3537	2662	3827	2944	2731	2842



41	ML000314a	4875	4087	4765	4996	3122	2269	2096	2356



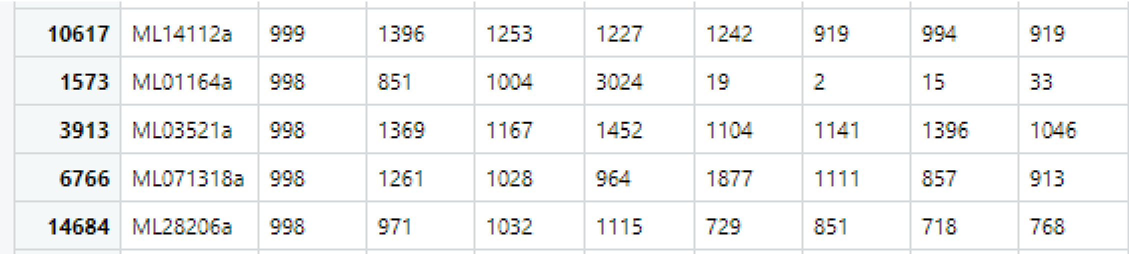
60	ML00062a	1857	1787	1972	1737	2901	1680	1332	1599

As you can see in the values above, although the values do fluctuate downwards in the V7, V8, V9 categories, the same is true for every single gene in the graph. And so if the case is that these 5 genes have the highest average expression, then that means that it will be consistent even if the column-based averages change

4.
The top 5 pairs of genes that I found to be closely related all had 0 expression. And so as a result there happened to be more than 5 pairs with this same level of correlation. I believe since there is 0 expression in the closely matching pairs of genes, this mean that they are close because they don't vary much

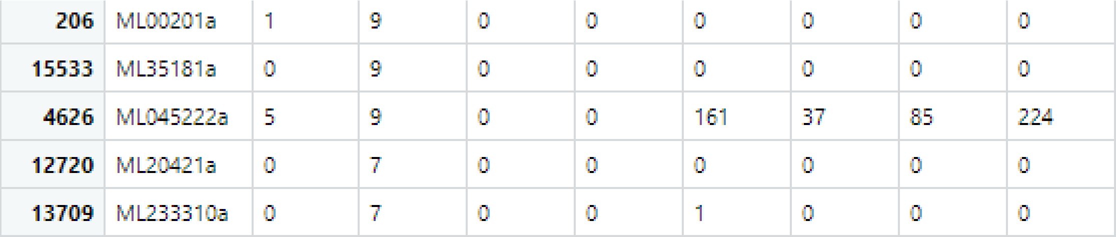
5.
The best way to divide genes in each column into high,medium and low count genes would be to use code that utilizes min, max data extraction, and one there are a certain amount of values left, these genes would be considered the values in the middle. The best way to describe this process would be to divide the data into thirds and remove the minimum values to obtain the low count genes, and maximum values to obtain the high count genes. What you are left with is a remaining third of the values that can be considered the medium count genes

6.
The top five genes with the most variability are as follows



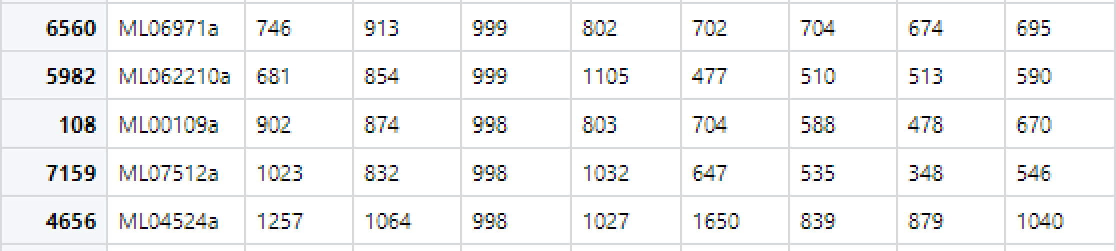
10617	ML14112a	999	1396	1253	1227	1242	919	994	919
1573	ML01164a	998	851	1004	3024	19	2	15	33
3913	ML03521a	998	1369	1167	1452	1104	1141	1396	1046
6766	ML071318a	998	1261	1028	964	1877	1111	857	913
14684	ML28206a	998	971	1032	1115	729	851	718	768

And these are the top five genes with the least variability



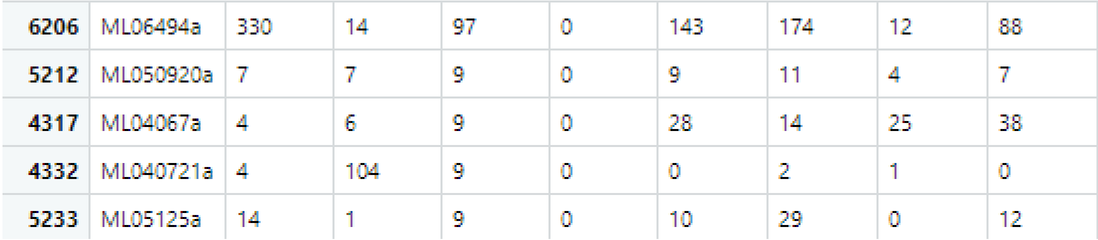
206	ML00201a	1	9	0	0	0	0	0	0
15533	ML35181a	0	9	0	0	0	0	0	0
4626	ML045222a	5	9	0	0	161	37	85	224
12720	ML20421a	0	7	0	0	0	0	0	0
13709	ML233310a	0	7	0	0	1	0	0	0

7. After using a t-test list the 5 most upregulated genes are



6560	ML06971a	746	913	999	802	702	704	674	695
5982	ML062210a	681	854	999	1105	477	510	513	590
108	ML00109a	902	874	998	803	704	588	478	670
7159	ML07512a	1023	832	998	1032	647	535	348	546
4656	ML04524a	1257	1064	998	1027	1650	839	879	1040

And the 5 most down regulated genes are



6206	ML06494a	330	14	97	0	143	174	12	88
5212	ML050920a	7	7	9	0	9	11	4	7
4317	ML04067a	4	6	9	0	28	14	25	38
4332	ML040721a	4	104	9	0	0	2	1	0
5233	ML05125a	14	1	9	0	10	29	0	12

If you rank by p value of the test, 1 or 2 tests change but the core 5 of the most up and down regulated still remain. I would not exclude these texts for having low expression because there are many genes that have 0 expression across the board. Compared to the genes with 0 expression, these down regulated genes have much higher values.