SPRING_18 STA_141A by Groupta Team Members: Hao Luo, Bianca Fung, Dai Chen

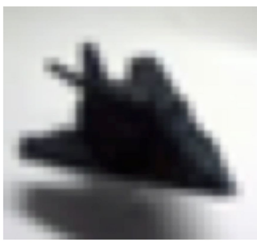
Qn3

Part(a) One image per class:

Total ten classes:

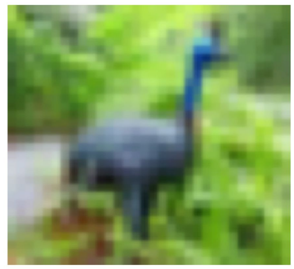
Plane, car, bird, cat, deer, dog, frog, horse, ship, and truck.

airplane

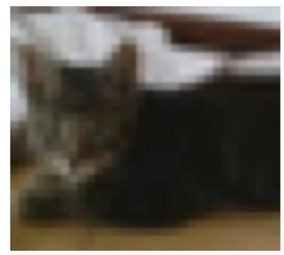


automobile

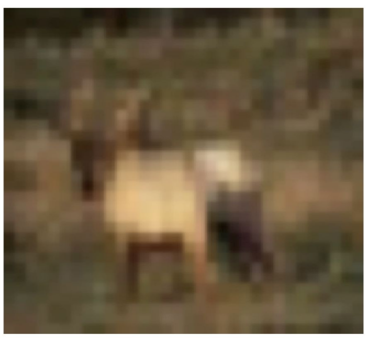
bird



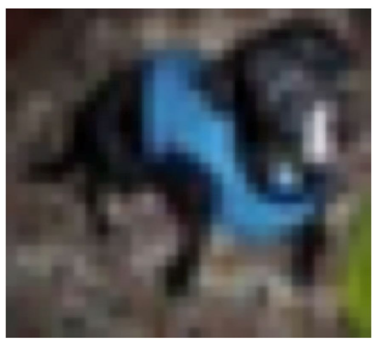
cat



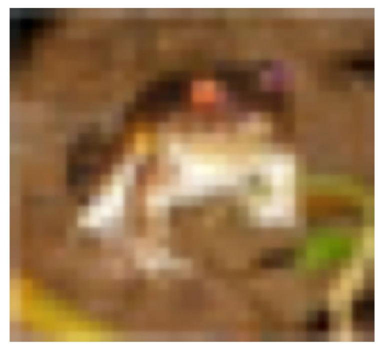
deer



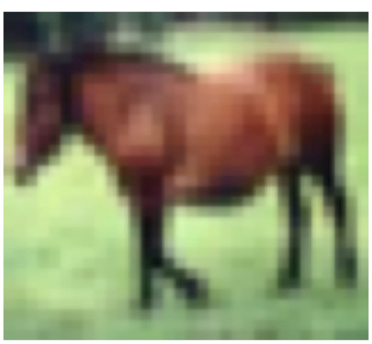
dog



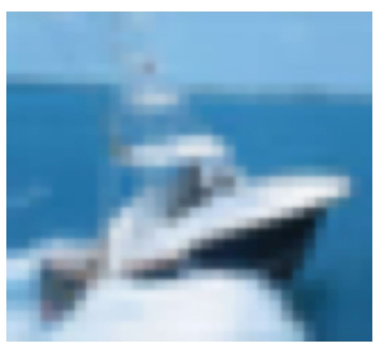
frog



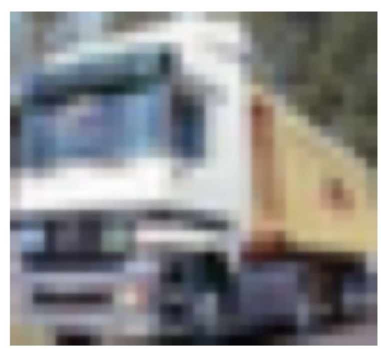
horse



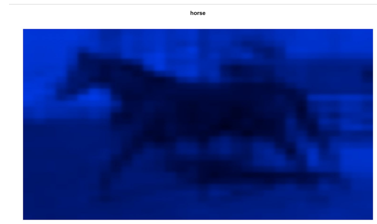
ship



truck



Part(b)Which color channels seem the most likely to be useful for classification?



Based on our discussion:

At first glimpse at the three original color pixel based photos. It seems that the green color is to be most useful because the contrast-ness on the horse and the background. However, computers see them differently and they read the pictures by numbers. Therefore, the pixels have the highest variation between classes but the least variation within the classes are good for classification (See the statistics below). As the result suggest, the most useful color channels for classification is BLUE and the least likely to be useful for classification is RED.

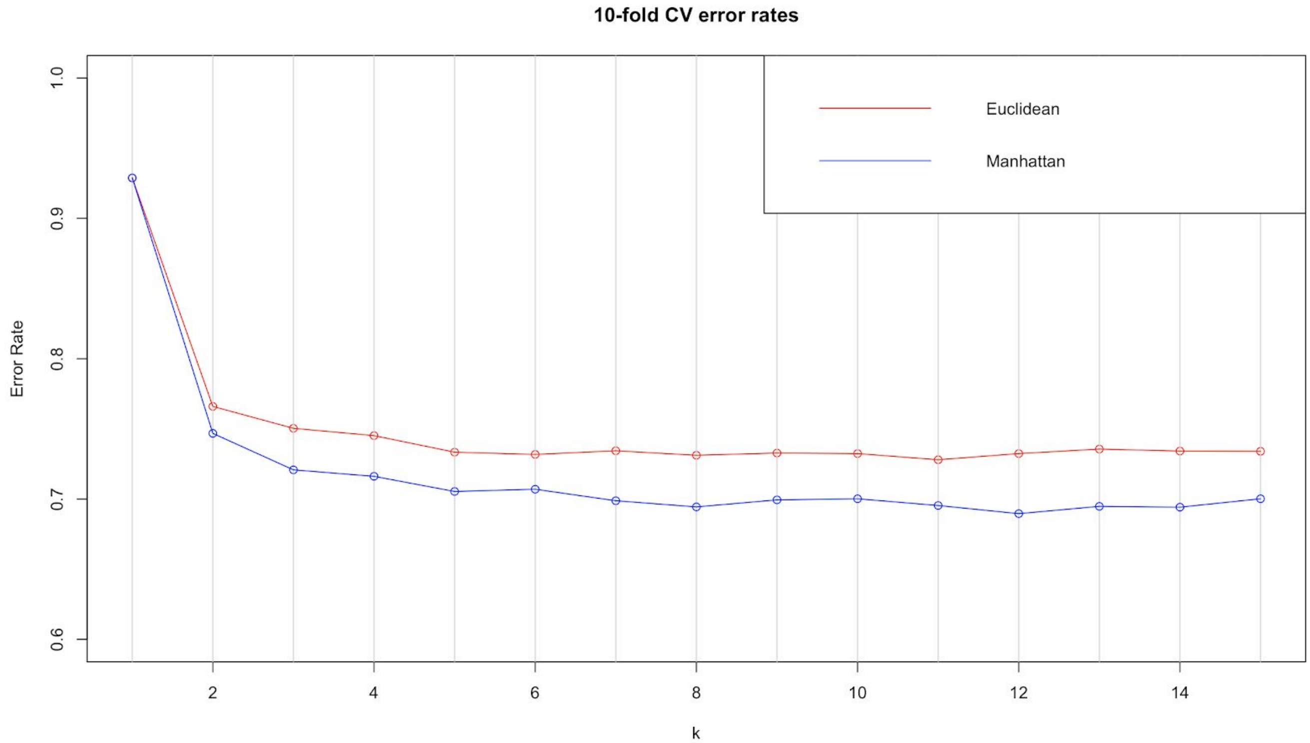
F*= msr/mse:

```
> sort(f_val, decreasing=TRUE)[1:10] # the top 10 highest F-value = most likely to be useful for classification
      B41       B7        B9       B39       B73       B13       B71       B11       B57       B25 
0.3129283 0.3120698 0.3107666 0.3106480 0.3090315 0.3073936 0.3064301 0.3062885 0.3058209 0.3051685 
> sort(f_val, decreasing=FALSE)[1:10] # the least 10 F-value = least likely to be useful for classification
       R572        R574        R542        R540        R604        R548        R554        R606        R538 
0.003561092 0.004434690 0.005029167 0.005212889 0.005460525 0.005958106 0.006027861 0.006042474 0.006343658 
       R506 
0.006517579 
```

Qn5

In order to shorten the run-time of the 10-fold cross validation, we perform the distance calculation at the beginning of the program. In particular, we allow user to input the calculated distance matrix. When we perform the 10-folds cross validation, instead of subset a fold and calculate the distance between one fold and the rest, we only need to use index to look up for the distance.

Qn6 & Qn7 & Qn8



As the statistics shows, that Manhattan method is better than Euclidean in this case.

From the confusion matrix below, we discovered that objects with clear square like shapes and features such as air planes , bird, deer, and ship are more likely to be correctly classified, which features provides more characteristics for the calculations. Therefore higher accuracy rate for the objects. However, due to the common features and shapes of ship and truck. Truck is likely to be classified as ship in both calculation methods.

On the other hand, objects with rounded like shapes and features, such as car , cat, dog, frog, and horse. From Manhattan method, the computer captures the main characteristics of the class but confused between them (see the Manhattan distance confusion matrix below). the They are less likely to be classified correctly. Moreover, color similarities and backgrounds of the photo are important factors too. Which will be discussed later.

For Euclidean Distance with k = 11

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	242	60	66	49	32	40	18	55	112	59
automobile	0	37	0	3	1	2	1	2	4	15
bird	63	78	210	134	144	112	157	123	28	69
cat	4	16	15	57	11	54	17	19	9	17
deer	37	120	145	130	249	151	155	184	49	63
dog	1	10	6	36	4	68	11	12	8	9
frog	23	34	24	59	28	46	132	26	8	36
horse	5	7	3	8	14	2	4	43	4	15
ship	125	118	29	23	16	24	5	33	272	167
truck	0	20	2	1	1	1	0	3	6	50

For Euclidean Distance with k = 8

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	240	71	63	45	37	47	21	57	133	70
automobile	0	43	2	2	2	1	1	3	4	20
bird	75	65	219	141	150	118	161	122	34	65
cat	5	18	21	63	18	59	23	22	10	18
deer	33	124	137	122	230	129	150	169	44	72
dog	1	12	3	42	5	71	9	13	11	11
frog	24	40	24	51	31	49	129	26	4	24
horse	5	5	3	9	11	5	3	50	5	13
ship	116	106	27	25	15	20	2	34	249	157
truck	1	16	1	0	1	1	1	4	6	50

For Euclidean Distance with k = 6

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	260	82	76	56	33	43	25	65	136	82
automobile	0	41	2	2	1	2	2	2	6	23
bird	60	65	207	131	155	115	158	112	31	67
cat	8	21	22	72	16	69	36	22	13	18
deer	37	121	129	121	229	128	147	166	42	60
dog	1	13	5	49	7	69	9	14	10	17
frog	21	35	27	41	32	45	115	27	7	25
horse	5	7	5	7	9	8	4	63	5	16
ship	108	100	26	20	17	19	3	24	243	150
truck	0	15	1	1	1	2	1	5	7	42

For Manhattan Distance with k = 12

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	260	66	70	51	44	45	28	62	136	82
automobile	1	88	1	2	1	5	4	4	8	36
bird	64	58	230	116	151	108	159	110	32	53
cat	3	24	22	82	16	65	27	25	5	26
deer	22	91	116	120	223	113	126	155	34	48
dog	3	8	6	36	5	88	7	14	8	10
frog	20	48	22	52	28	44	141	16	9	24
horse	7	13	4	15	14	8	4	79	3	27
ship	118	82	24	24	15	22	3	26	256	113
truck	2	22	5	2	3	2	1	9	9	81

For Manhattan Distance with k = 14

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	255	48	68	45	40	39	25	51	106	68
automobile	3	69	0	5	1	8	3	0	5	28
bird	56	69	225	120	147	111	157	109	27	54
cat	7	18	20	64	8	48	17	21	6	14
deer	21	108	115	117	237	126	144	169	42	42
dog	1	8	9	35	4	89	12	11	7	5
frog	22	42	22	68	31	43	130	22	6	34
horse	4	7	6	13	11	8	6	78	5	23
ship	128	98	32	27	18	28	5	28	287	137
truck	3	33	3	6	3	0	1	11	9	95

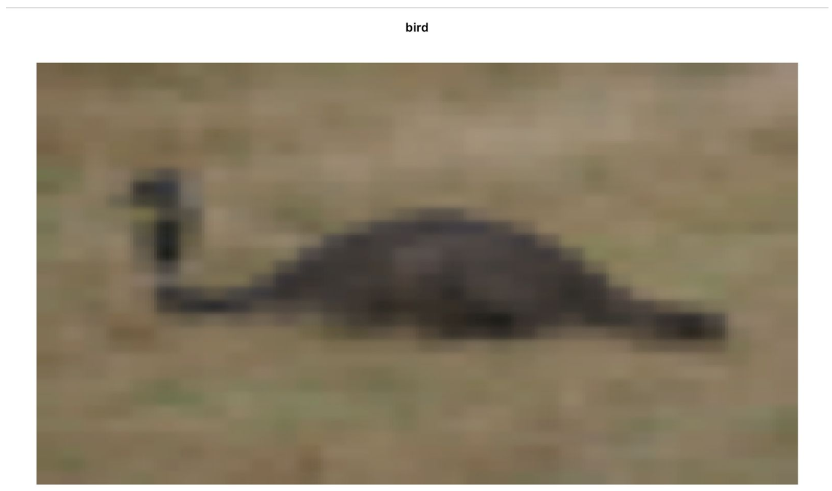
For Manhattan Distance with k = 8

predict \ true	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	260	66	70	51	44	45	28	62	136	82
automobile	1	88	1	2	1	5	4	4	8	36
bird	64	58	230	116	151	108	159	110	32	53
cat	3	24	22	82	16	65	27	25	5	26
deer	22	91	116	120	223	113	126	155	34	48
dog	3	8	6	36	5	88	7	14	8	10
frog	20	48	22	52	28	44	141	16	9	24
horse	7	13	4	15	14	8	4	79	3	27
ship	118	82	24	24	15	22	3	26	256	113
truck	2	22	5	2	3	2	1	9	9	81

The 3 best k is shown as below:

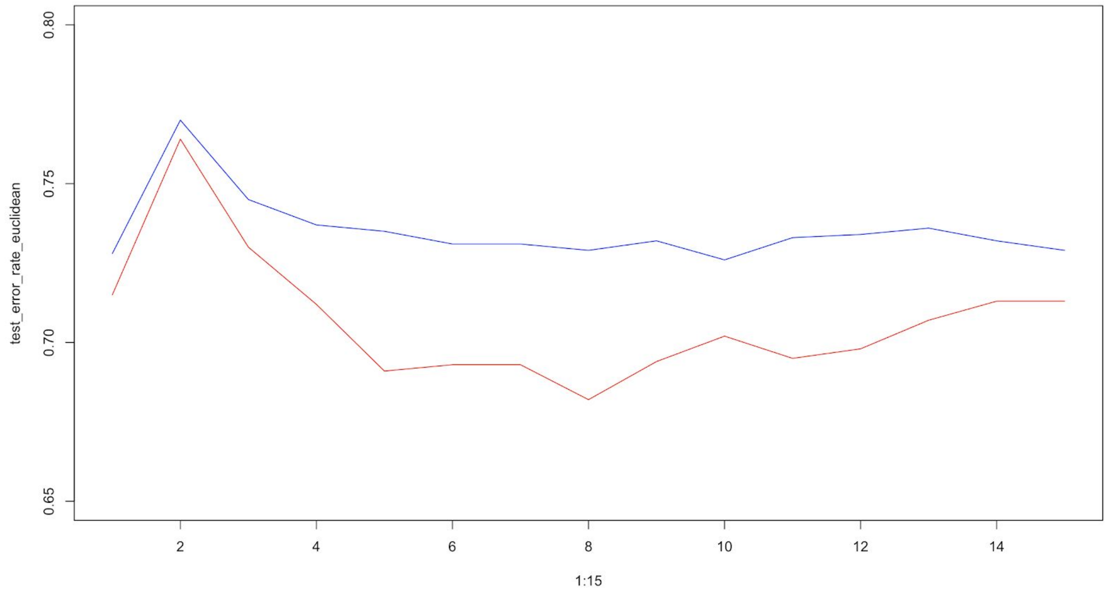
```
> # overall misclassification rate
> euc_acc = c(sum(diag(euc_conf1)), sum(diag(euc_conf2)), sum(diag(euc_conf3)))/5000
> man_acc = c(sum(diag(man_conf1)), sum(diag(man_conf2)), sum(diag(man_conf3)))/5000
> euc_acc # 0.2720 0.2688 0.2682
[1] 0.2720 0.2688 0.2682
> man_acc # 0.3056 0.3058 0.3056
[1] 0.3056 0.3058 0.3056
> # best combinations:
> # 1: (man, k=14)
> # 2: (man, k=12)
> # 3: (man, k=8)
```

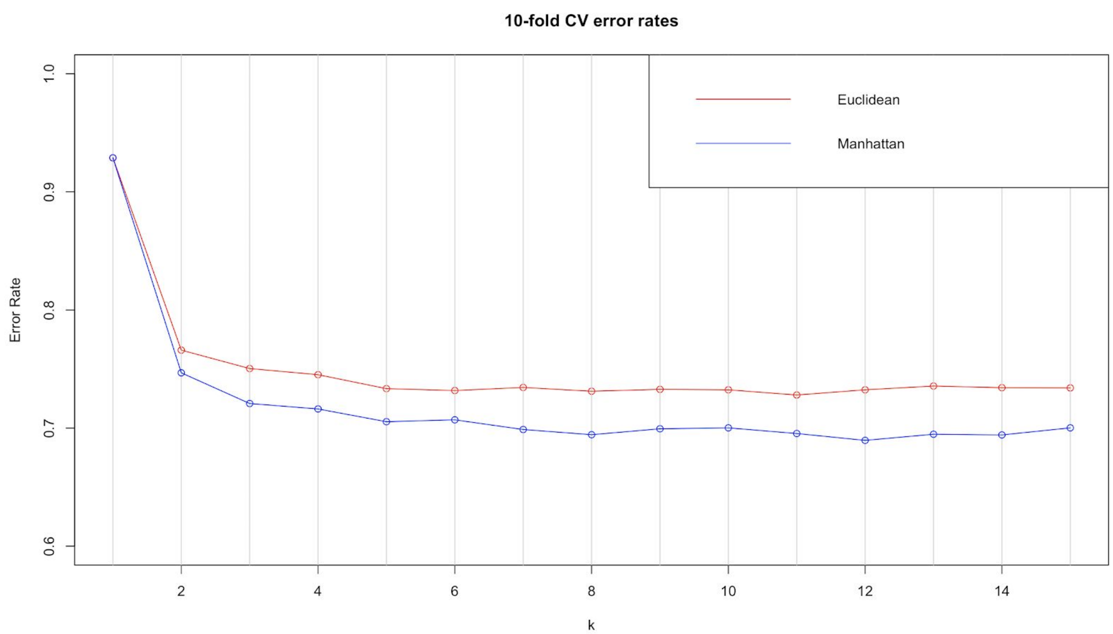
Discussion on misclassifications:



The pictures above were frog and bird. Due to the lack of training data sets and the similarities of the two. The bird on the right was misclassified as the frog on the left. Moreover, the overall colors and shapes share some characteristics. It makes senses that there were confusions as the their distance are close to each other.

Qn9





Comparing Manhattan to Euclidean in this case, Manhattan has more advantages in this case as it considers less outliers than Euclidean. Whereas Euclidean considers more data and it is more spread out than Manhattan. If the data set is large enough Euclidean should have a better result.

QN 10:

We collaborate and discuss the questions and approaches to each one of them.

Hao Luo is responsible for coding and assembling information.

Bianca Fung and Dai Chen are responsible for researching and finding related information regarding to the problems, and bring them to discussion during group meets.