James Yang

DATA 558

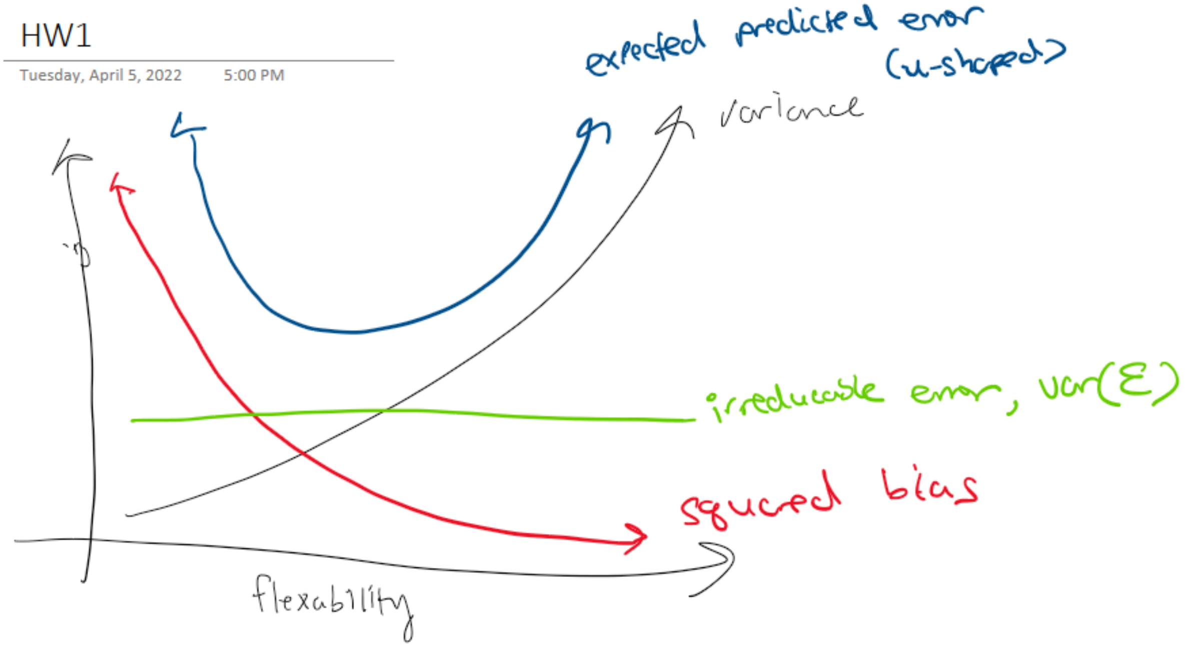
Statistical ML

Homework #1

1. Suppose that you are interested in performing regression on a particular dataset, in order to answer a particular scientific question. You need to decide whether to take a parametric or a non-parametric approach.
    a. A pro of a parametric approach is assuming a form of the function as a model. It simplifies the scientific problem by making it easier to estimate values within the model (i.e β0,β1, . . . ,βp). A con of a parametric approach is that the model chosen will not usually match the true unknown form of the function. Almost nothing in life will be a perfect discriminant in an LDA model because there will almost always be discrepancies.

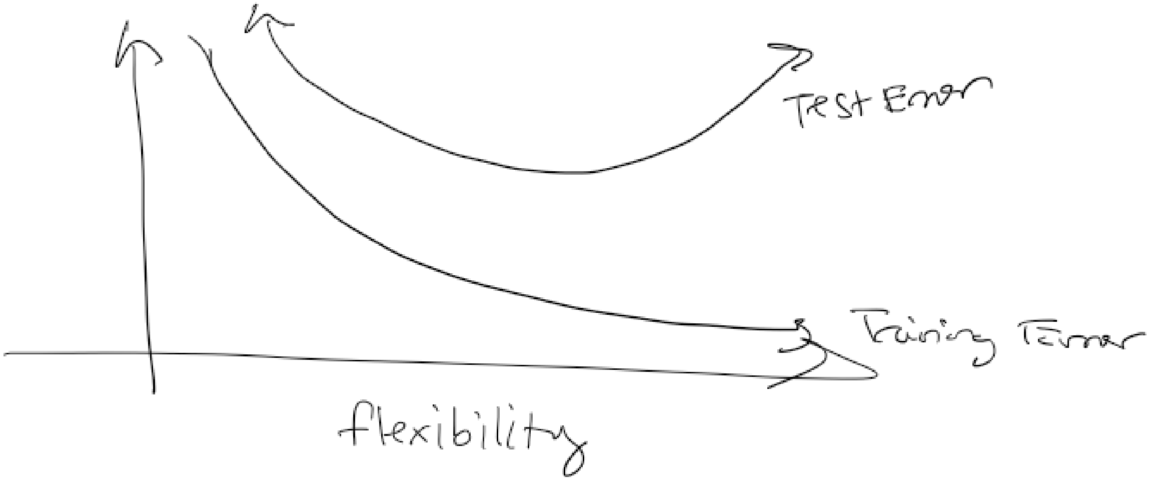
       A pro of a non-parametric approach is not having to succumb to a pre-determined form of a function. It gives the potential of accurately fitting a wider range of possible shapes for the function. A con of a non-parametric approach is not being able to reduce the problem of estimating the function to smaller parameters. A large number of observations are typically needed to obtain an accurate estimate of the function.
    b. Surveying the dataset and finding out the distribution on a histogram is a step in determining whether it is symmetrical and normally distributed. If it is normally distributed, a parametric approach would be ideal.
    c. When surveying the dataset, if you find that the distribution on a histogram isn't relatively normal (even after removing outliers and such in the data), then we can assume that a non-parametric approach would be ideal.
2. In each setting, would you generally expect a flexible or an inflexible statistical machine learning method to perform better? Justify your answer.
    a. When the sample size is small and the predictors are large, **an inflexible method** would perform better because if we use a small number of sample sizes, a flexible method would overfit the data, meaning they follow the errors, or noise too closely.
    b. When the sample size is large and the predictors are small, **the flexible method** would do better because it'll fit the data closer to what it needs to be. Since the sample size is large, it would also perform better than an inflexible approach because flexible models require estimating a greater number of parameters.
    c. When the predictors and response is highly non-linear, **a flexible method** would perform better because we want something with more freedom in morphing into the function than a non-flexible method. Trying to fit a linear model onto a multimodal logistic distribution is terrible.
    d. If the variance of the error in the data is high, we would want an **inflexible method** to eliminate the outstanding errors. If we were to incorporate a flexible method onto this model, it would include more error terms and make the model even more wrong. This would also increase its variance even more.

3. For each scenario, determine whether it is a regression or a classification problem, determine whether the goal is inference or prediction, and state the values of n (sample size) and p (number of predictors).
    a. This problem is a regression problem. The goal is a prediction problem and the value of n = 50 and p = 8.
    b. This problem is a classification problem. The goal is an inference problem and the value of n = 50 and p = 6.
4. This problem has to do with the bias-variance trade-o_ and related ideas, in the context of regression. For (a) and (b), it's okay to submit hand-sketched plots: you are not supposed to compute the quantities referred to below on data; instead, this is a thought exercise.
    a.



Note: Reducible Error = Bias^2 + Variance.
The level of flexibility considered best is the variance and the expected predicted error. As the number of samples increase, the variance flexibility will increase due to the error that it will eventually accumulate. However, the expected predicted error will increase as the noise increases.

    b. The test error has the best flexibility.

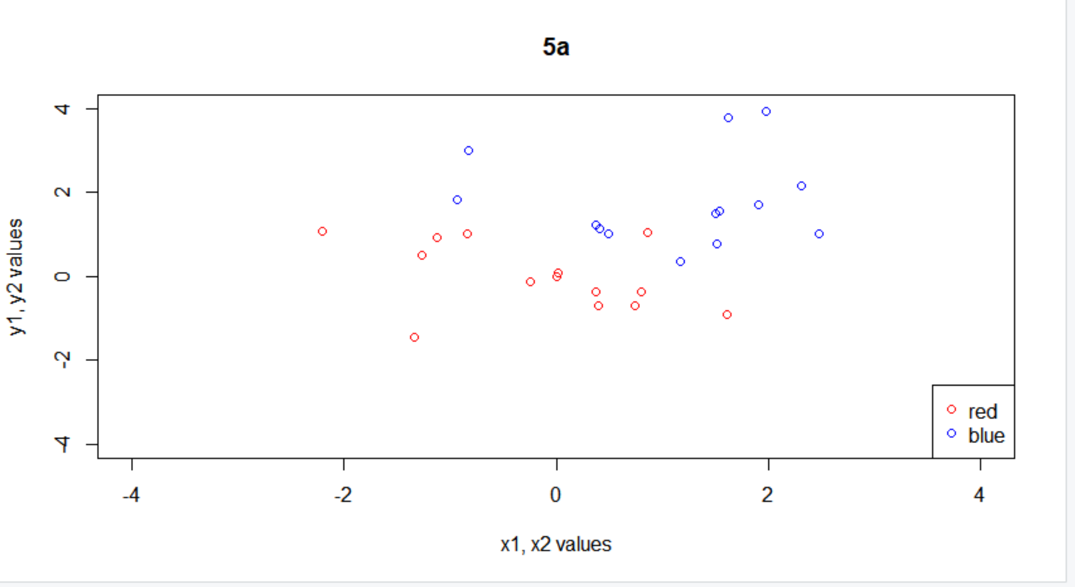


c. A function that has low bias and high variance is a **decision tree**. For example, an algorithm used to predict the weather forecast. If it takes a path of sunny, we need to check if it is cold or warm and so on. This algorithm doesn't take bias because it is simply stating what is going on with the weather, and it is extremely high in variance because the weather could be cloudy, gloomy, rainy, sunny, etc.

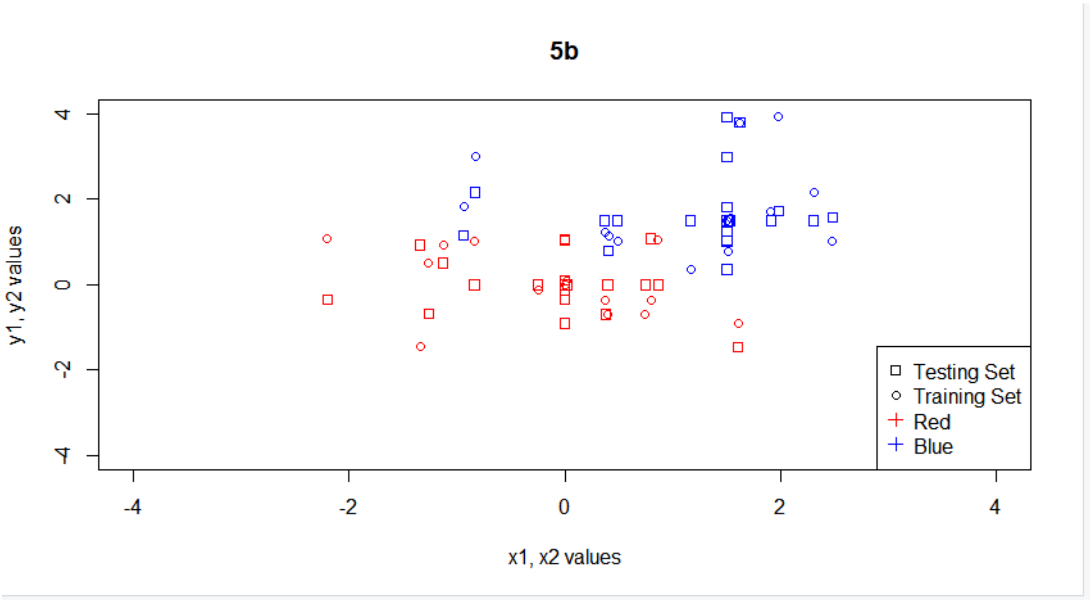
d. A function that has high bias and almost no variance would be **a linear regression algorithm**. This could be used to predict future values of the temperature by giving it a strict value following a trend line. The bias could be incredibly high based on the sample that it regresses from, and how the datapoints were retrieved. It could be underfitted and many other factors.

5.

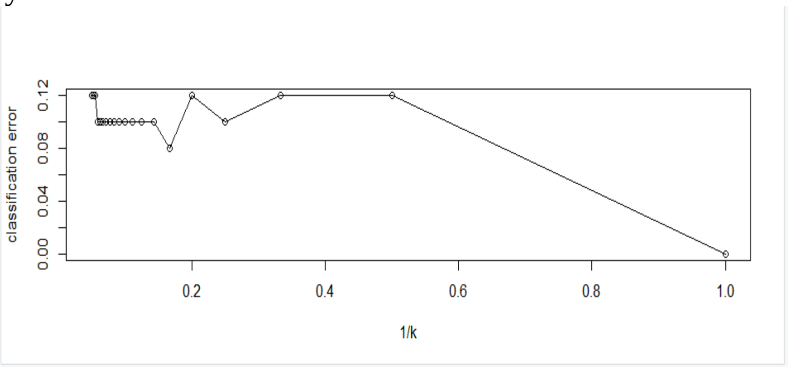
a. The following is my graph.



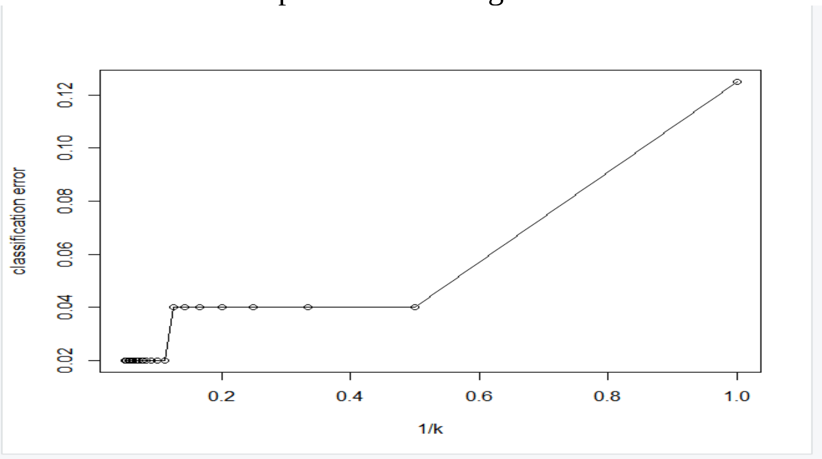
b.



c. Using the Knn function, my errors come out to about this for Testing and Training respectively:



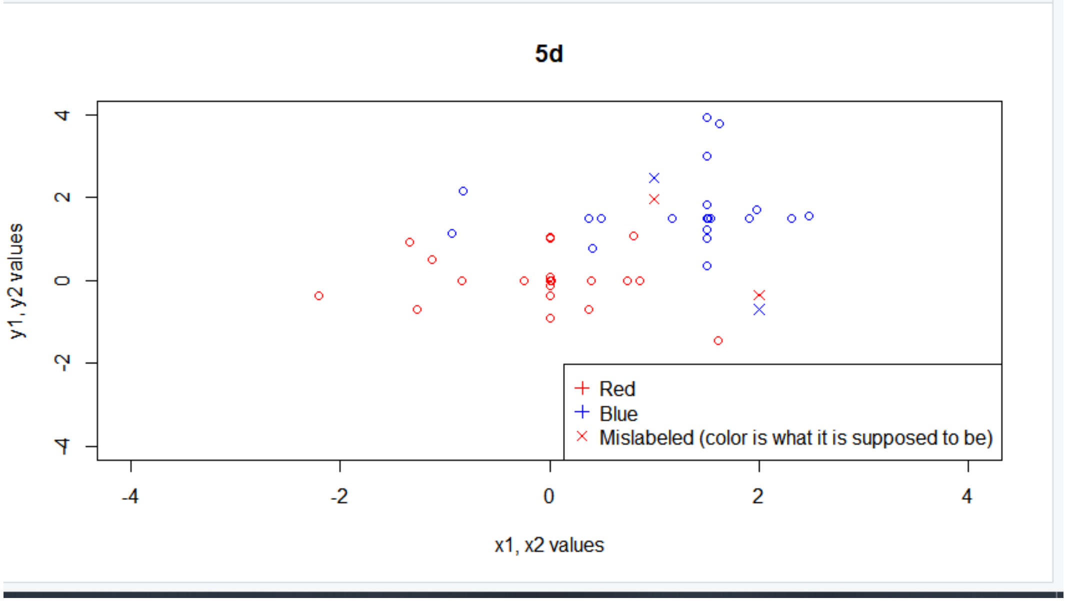
Depicts the Training KNN



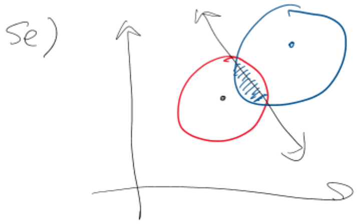
Depicts the Testing KNN

This graph makes sense because the error rate at K=1 is always going to be zero for a training sample because any closest point to that point is basically itself. The best classification error varied from **K = 10 to K = 6 interestingly**. It would also fluctuate from **0.08 to 0.04** error rate when simulating. I chose the higher value to be safe. When the error rate approaches the size of the training set, it will surely increase. 6 is a bit far from the size of the training set, which also explains why it isn't at the highest error percentage of 12%.

d. For K= 6, error rate of 8% (4 elements)



e. In class, we discussed how to solve this problem through sampling.



Red : $N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$

Blue: $N\left(\begin{pmatrix}1.5\\1.5\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$

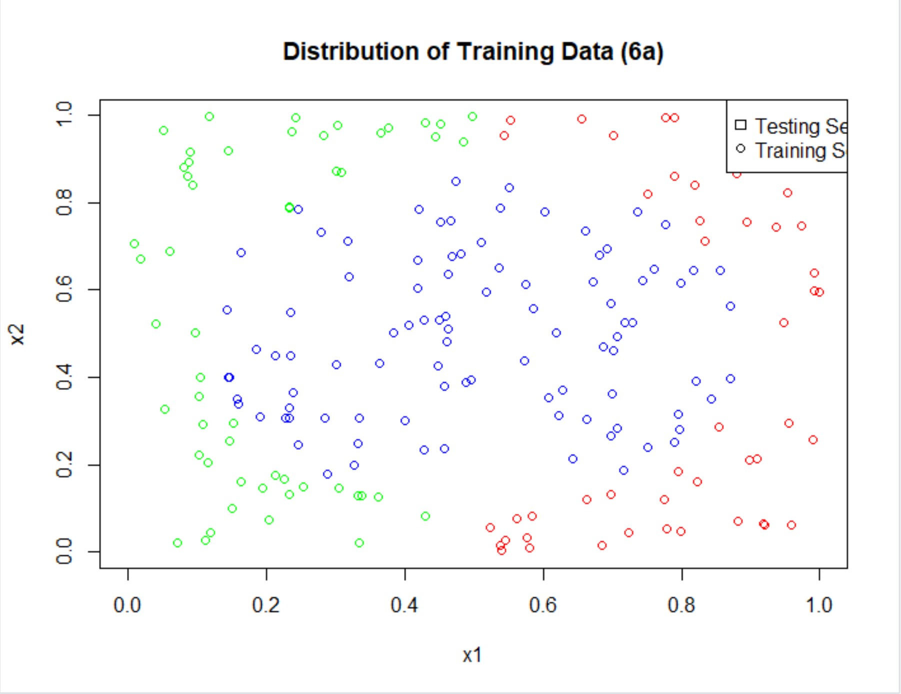
$x+y=1.5$

To find error rate,
sample to get percentage
of how many.

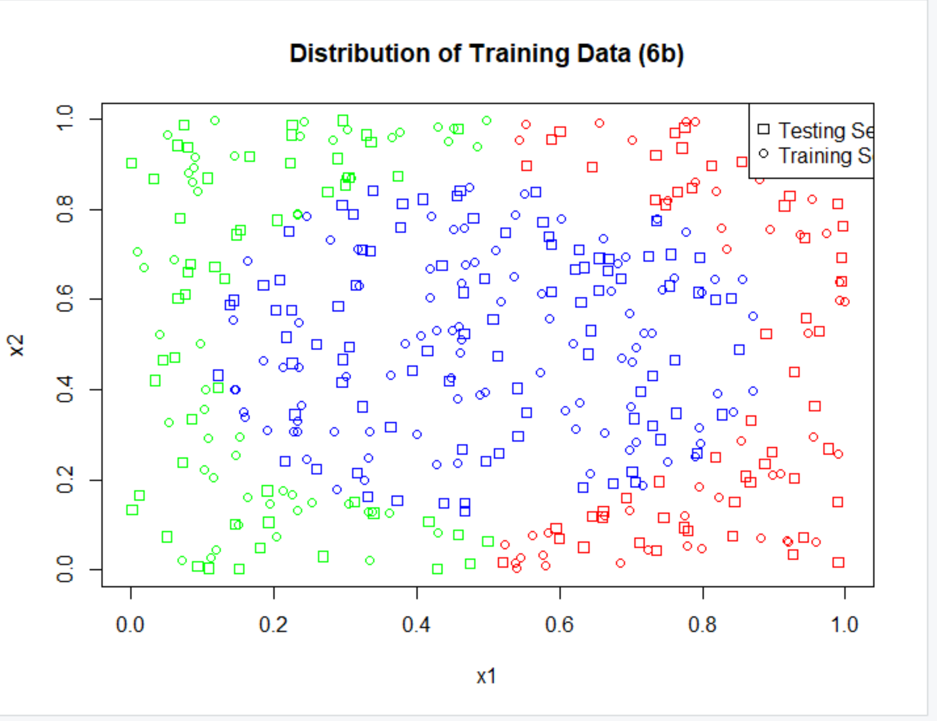
After sampling 10,000 values of this distribution using a line of x+y=1.5, I found that approximately 1544 values were incorrectly labeled, giving it a **15.4% error rate.**

6.

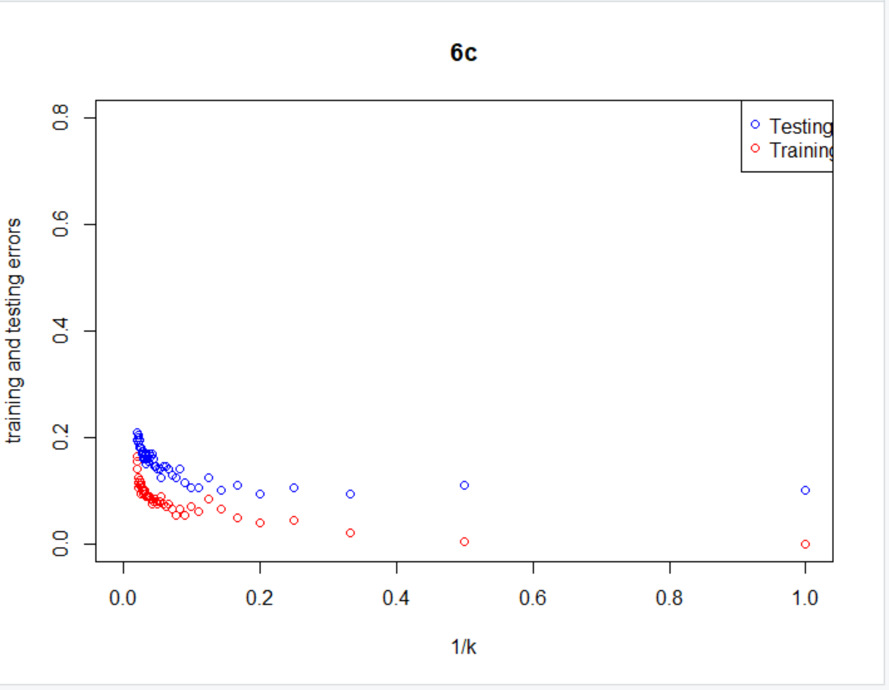
a.



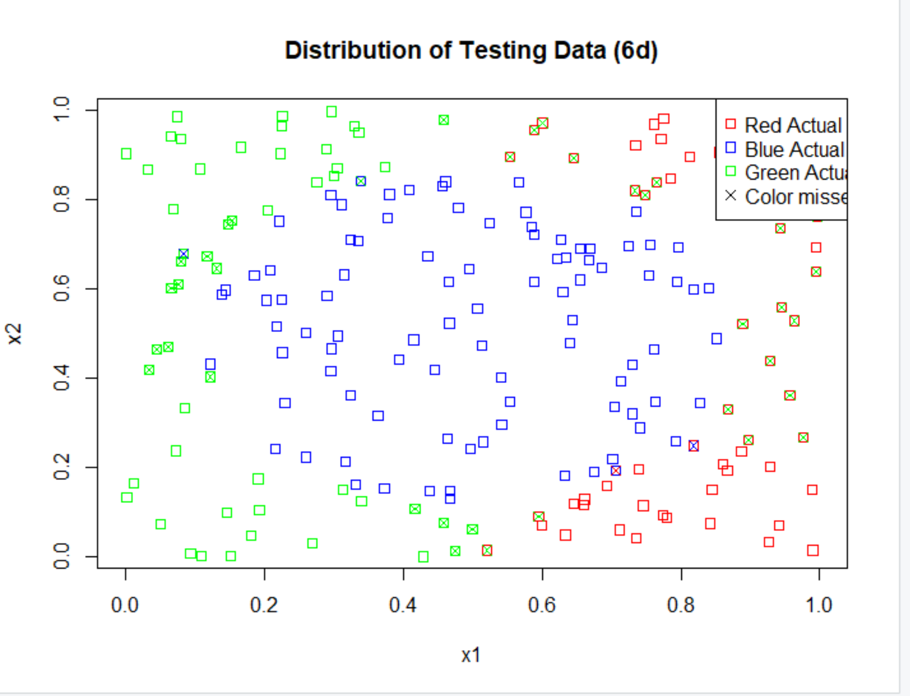
b.



c. This answer makes sense because it actually resembles a reflection of the graph that we created in the earlier section with the testing and training graphs. The classification error is less when K gets higher because this model has a lot higher change in plotting points. The last question only handled two, but this time it is concurrent with 3 distributions creating a greater variance. Obviously when K = 1, the training set gets set to 0 because it'll just find the actual values next to it.



d. For the value of **K = 37,** the model performed best with a 0.095 classification error

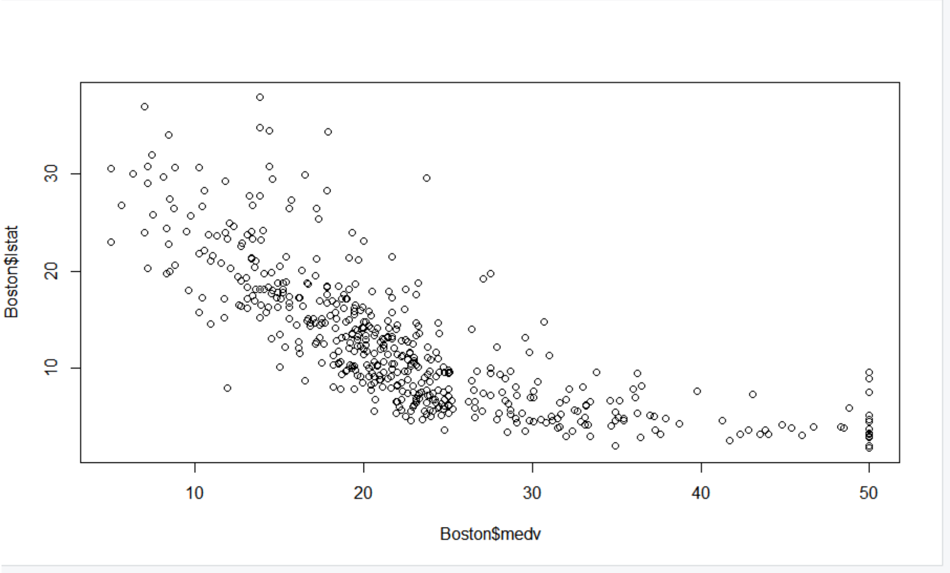


e. The error rate was approximately 24% from testing a bunch of rapid samples. About 152 predictions were correctly predicted and the reason behind this is because Bayes is a linear classifier while KNN is not. What this means is that unlike question 5, we cannot apply a linear line to determine which color is which. This correlates well with its higher error rating in comparison to question 5.

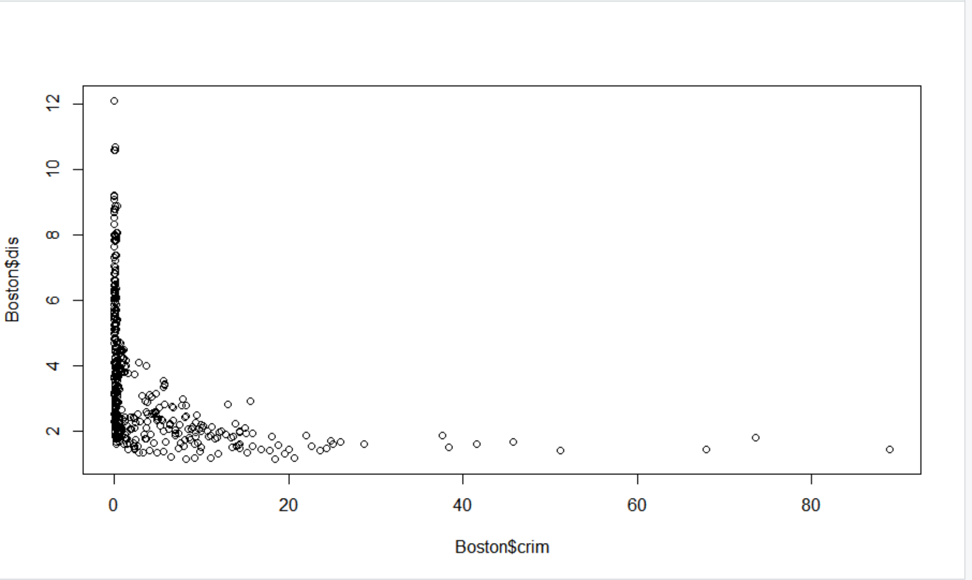
7.

a. There are 506 rows and 13 columns. The data set contains housing values in 506 suburbs of Boston. Crim represents the crime rate per capita by town. Zn is the proportion of residential land zoned for lots over 25,000 sq ft. Indus is the proportion of non-retail business acres per town. Chas is the Charles River dummy variable. Nox is the nitrogen oxides concentration. Rm is the average number of rooms per dwelling. Age is the proportion of owner-occupied units built prior to 1940. Dis is the weighted mean of distances to five Boston employment centers. Rad is the index of accessibility to radial highways. Tax is the full-value property-tax rate per $10,000. Ptratio is the pupil-teacher ratio by town. Lstat is the lower status of the population. Medv is the median value of owner-occupied homes in the $1000s.

b. It appears that when plotting a pairwise of the lower statistical income people and their housing (in the $1000s), there are a lot lower income people owning houses in the lower median cost of owner-occupied homes. There is obvious correlation in these two variables and it makes sense as well. There are outliers but they do not represent the bulk of the dataset.



When plotting the distance between 5 employment centers and crime rate, it appears there is little correlation between these two ideas as they appear to have little impact on each other besides a few outliers in the dataset which may be coincidental.



c. Interestingly there is a correlation between a few of the predictors. Nox, Rm, Lstat, and medv appear to have correlation in the body of the dataset. A lot of these predictors appear to have outliers that do not correlate, but their main body having interesting correlation.

d. .

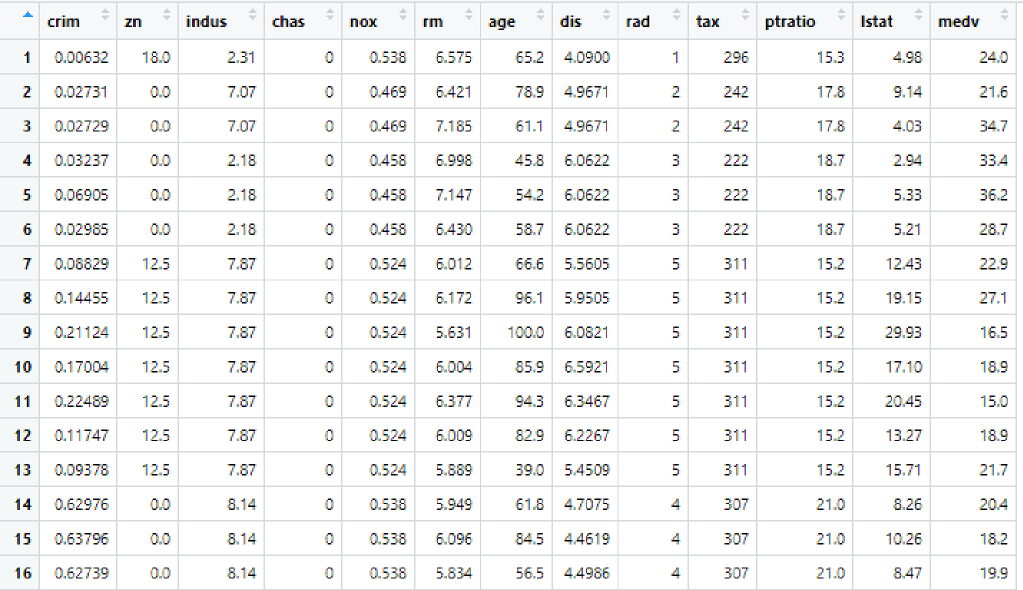
Predictors	Min	Max
Crim	0.00632	88.9762
Zn	0.00000	100.0000
Indus	0.46000	27.7400
Chas	0.00000	1.0000
Nox	0.38500	0.8710
Rm	3.56100	8.7800
Age	2.90000	100.0000
Dis	1.12960	12.1265
Rad	1.00000	24.0000
Tax	187.00000	711.0000
Ptratio	12.60000	22.0000
Lstat	1.73000	37.9700
Medv	5.00000	50.0000

The one thing I have noticed is the abnormally high tax rate to median house value. It seems a bit odd how the median tax value could go up to 7.1 million when the top house value is only worth $50,000. I'm not sure if that tax rate included other things such as business, but it seems like an odd outlier.

e. 35 suburbs have a bound to the Charles River.

f. Pupil-teacher ratio mean = 18.45
Pupil-teacher ratio SD = 2.164

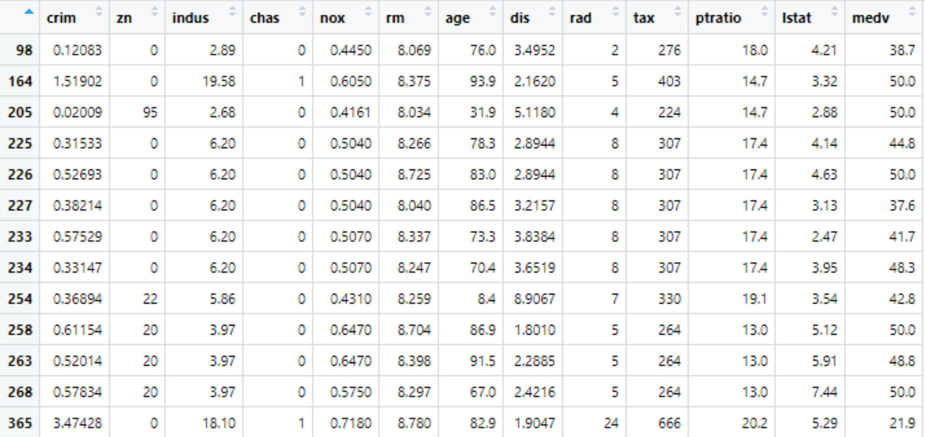
g. Out of the suburbs in Boston, the highest median value of owner-occupied homes is $50,000 and 16 suburbs have that value.



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0
2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4
5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2
6	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	17.10	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	20.45	15.0
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	13.27	18.9
13	0.09378	12.5	7.87	0	0.524	5.889	39.0	5.4509	5	311	15.2	15.71	21.7
14	0.62976	0.0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21.0	8.26	20.4
15	0.63796	0.0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21.0	10.26	18.2
16	0.62739	0.0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21.0	8.47	19.9

Here are the values of all suburbs that had the top median values. The values are a bit scattered on the range for most values. It's interesting how half of the values contain a Zn and the other half don't have any at all. Something else that is worth noting is that the tax range is all in the high 200s to low 300s, making them very close to the median value of 330. They are also in the median value of the ptratio. The business ratios found around these homes were not very high, implying that they were around more residential neighborhoods rather than stores. We can pull from the information that parts of the surveyed neighborhoods have industries around them while the other half do not, giving tax values that are similar to each other but in totally different areas of Boston. They each have their relative lower income status, most of which relating more to the areas that have more crime.

h. 13 suburbs average more than eight rooms per dwelling and 333 suburbs average more than 6 rooms.



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
98	0.12083	0	2.89	0	0.4450	8.069	76.0	3.4952	2	276	18.0	4.21	38.7
164	1.51902	0	19.58	1	0.6050	8.375	93.9	2.1620	5	403	14.7	3.32	50.0
205	0.02009	95	2.68	0	0.4161	8.034	31.9	5.1180	4	224	14.7	2.88	50.0
225	0.31533	0	6.20	0	0.5040	8.266	78.3	2.8944	8	307	17.4	4.14	44.8
226	0.52693	0	6.20	0	0.5040	8.725	83.0	2.8944	8	307	17.4	4.63	50.0
227	0.38214	0	6.20	0	0.5040	8.040	86.5	3.2157	8	307	17.4	3.13	37.6
233	0.57529	0	6.20	0	0.5070	8.337	73.3	3.8384	8	307	17.4	2.47	41.7
234	0.33147	0	6.20	0	0.5070	8.247	70.4	3.6519	8	307	17.4	3.95	48.3
254	0.36894	22	5.86	0	0.4310	8.259	8.4	8.9067	7	330	19.1	3.54	42.8
258	0.61154	20	3.97	0	0.6470	8.704	86.9	1.8010	5	264	13.0	5.12	50.0
263	0.52014	20	3.97	0	0.6470	8.398	91.5	2.2885	5	264	13.0	5.91	48.8
268	0.57834	20	3.97	0	0.5750	8.297	67.0	2.4216	5	264	13.0	7.44	50.0
365	3.47428	0	18.10	1	0.7180	8.780	82.9	1.9047	24	666	20.2	5.29	21.9

Something I notice immediately is that the crime rate is mostly in the same range (0.3-0.6) except for a few outliers. These suburbs are on the older side (71), and they all have approximately a similar ptratio. Their median value of house is also quite similar (on the higher end). Houses in Boston that have 8 or more rooms per dwelling are typically an older home in a residential area with a similar crime rate and a higher price range.