1.
    a. A flexible method would work better, as it has a large sample size and thus meets the requirements of flexible methods to have large n. Further, the number of predictors are smaller and so will generally be more flexible.
    b. Inflexible methods would be better, because n is small and too many predictors can cause flexible methods to be overfitted.
    c. Flexible methods would be best because inflexible methods do not perform as well for non-linear relationships.
2.
    a. This is a regression problem interested in inference. N = 500, p = 3
    b. This is a classification problem interested in prediction. N = 20, p = 13
    c. This is a regression problem interested in prediction. N = 52, p = 3
4.
    a. Classification
        i. A response may be whether a skin mole is malignant or benign, and the predictor variables would be items such as diameter, color, border pattern. In this case, we would be more interested in which variables contribute to malignancy and thus the goal would be inference.
        ii. We may be interested in predicting whether a patient will contract HIV, so the outcome would be HIV status, with predictors such as whether the patient is associated with an HIV cluster, behavioral factors such as their engagement in risky behaviors such as intravenous drug use or unsafe sex practices, race, factors relating to economic status, and years of education. As the overall goal would be to determine future HIV status, this is an example of a prediction application.
        iii. We may be interested in whether a student received a medical school offer. Predictor variables may include college GPA, MCAT score, hours of volunteering, hours of shadowing, and whether someone in their family was a doctor. This is an example of a prediction application, as we are more interested in determining whether a student will be accepted.
    b. Regression
        i. We may be interested in ascertaining college success (measured by college GPA) using variables such as high school GPA, involvement in extracurriculars, and SAT/ACT scores. This could be an example of prediction, as we are more interested in their ending performance than what contributed towards good or bad performance.
        ii. An outcome variable may be birth weight with predictor variables such as mother's smoking status, weight, and use of estrogen products. In this situation, we are interested in which factors are related to low birthweight, so it is an example of an inference application.
        iii. We may be interested in what factors lead to high poverty levels. By examining factors such as educational budget, unemployment, crime rates,

and social support indexes in various communities, we may be able to *inference* which factors most contribute to high poverty.

c. Cluster Analysis
    i. We may be interested in grouping flower species together to find evolutionarily linked species based on sepal and petal length and width.
    ii. Examining factors such as age, sex, and spending habits may help businesses identify groups and subgroups with commonalities that they can target.
    iii. Insurance companies may use cluster analysis to identify groups of people who are more high risk (and should have higher rates) versus people who are low risk (and should have lower rates).

5. Flexible approaches generally decrease bias but increase variance, and work particularly well in situations where a linear approximation is not appropriate and we have a large set of observations. Inflexible approaches can introduce more bias, but have less variance. They work well in situations where a linear model is reasonable or we do not have many observations. Inflexible models are also generally more interpretable, so they are better when are interested in inference. However, when focused mostly on prediction, a flexible method may be the best method.

7.

X1	X2	X3	Y	distance
0	3	0	red	3.000000
2	0	0	red	2.000000
0	1	3	red	3.162278
0	1	2	green	2.236068
-1	0	1	green	1.414214
1	1	1	red	1.732051

a. The Euclidean distance from the origin for each observation is in the "distance" column.
b. The test point will be green, as the nearest point to the origin is (-1,0,1) which is green.
c. The test point will be red, as the majority closest to the origin (test point) are red [(2,0,0) and (1,1,1)] with one point being green [(-1,0,1)].
d. We would expect the best K to be smaller, because as K grows it becomes less flexible and more linear.

8.

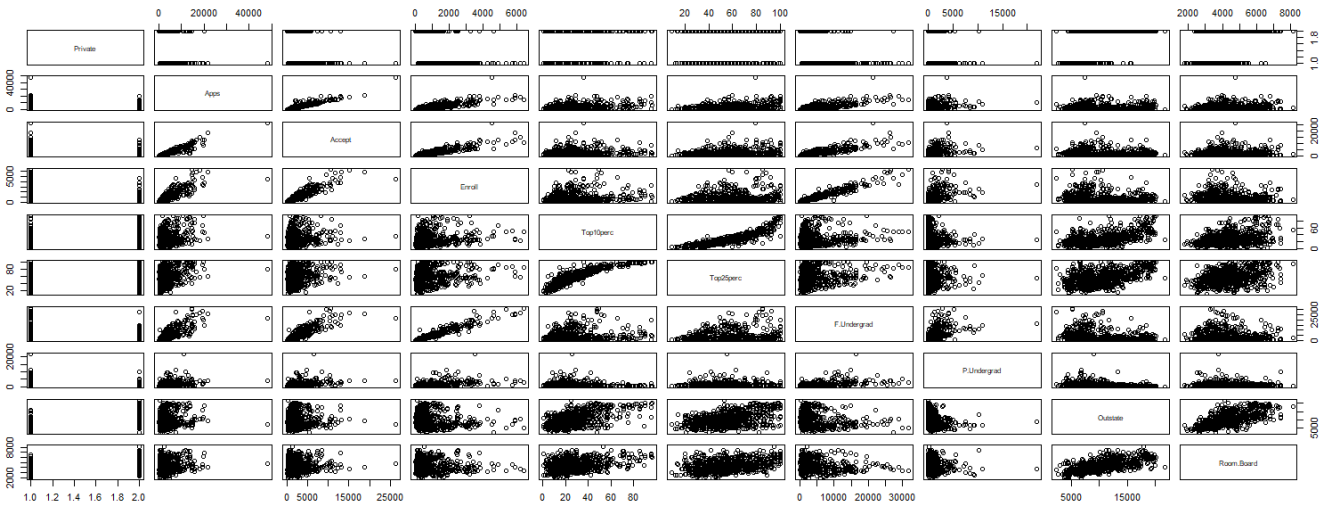
c.

i. The maximums of a few variables indicate there may be issues with data collection. For example, the percentage of faculty with PhD has a maximum of 103%; similarly the maximum of graduation rate is

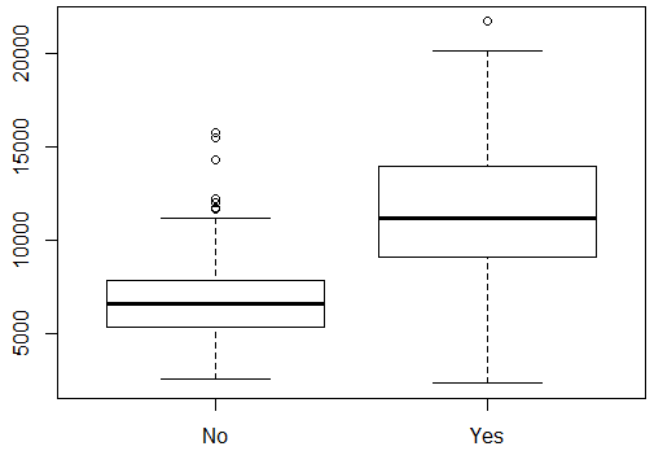
```
> summary(college)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc      F.Undergrad
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 1707
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
  P.Undergrad         Outstate       Room.Board       Books           Personal         PhD          Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

118%. This could be a misinterpretation of units, but makes interpreting the data difficult. It is also easy to identify outliers within certain categories. For example, the mean and median number of applications are 1558 and 3002, indicating skew but the maximum of 48, 094 appears to be a fairly extreme outlier. Similarly, the median cost of books is $500, but one school states a cost of $2340, which is much higher than the 3rd quartile value of $600.

ii. There are quite a few strong, seemingly linear associations in this dataset. As mentioned above, there is one far outlier in the number of applications. The number of applications has a positive association with the number of acceptances, enrolled, and full time undergraduates. Similarly, the number of acceptances is positively associated with full time undergraduates as well. Enrollment is also appears to be positively associated with full time undergraduates. Unsurprisingly, there appears to be a strong association between the number of new students in the top 10% and top 25% of their high school class. There also appears to be a positive association between room and board costs and out-of-state tuition.

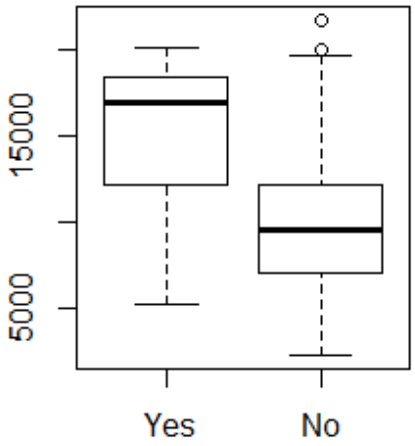


iii. Non-private schools have more outliers but less variance, and have a median out-of-state tuition of around $6000. Private schools have more variance with a median of about $10,000 and a maximum over $20,000.

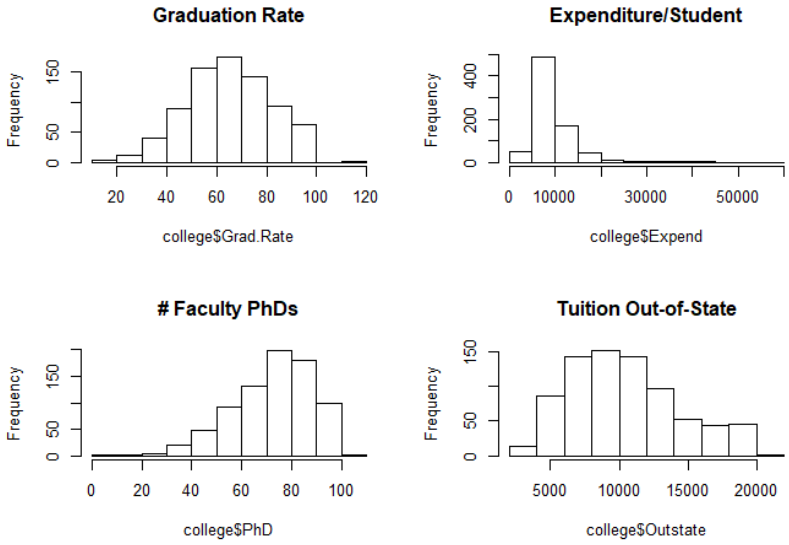


iv. There are 78 elite schools and 699 non-elite schools in this dataset. Elite schools have a median out-of-state tuition of about $16,000 while non-elite schools have a median out-of-state tuition of about $10,000. Non-elite schools have a slightly larger variance with more high outliers and a higher maximum.

```
> summary(college$Elite)
 Yes   No
  78  699
>
```



v. Most schools have a mean graduation rate of about 60%. Most schools' faculty comprises of at least 60% PhDs. Schools generally spend $1000 – 2000 in expenditures per student, while out-of-state tuition costs largely vary but with most schools charging between $7,500– $15,000.

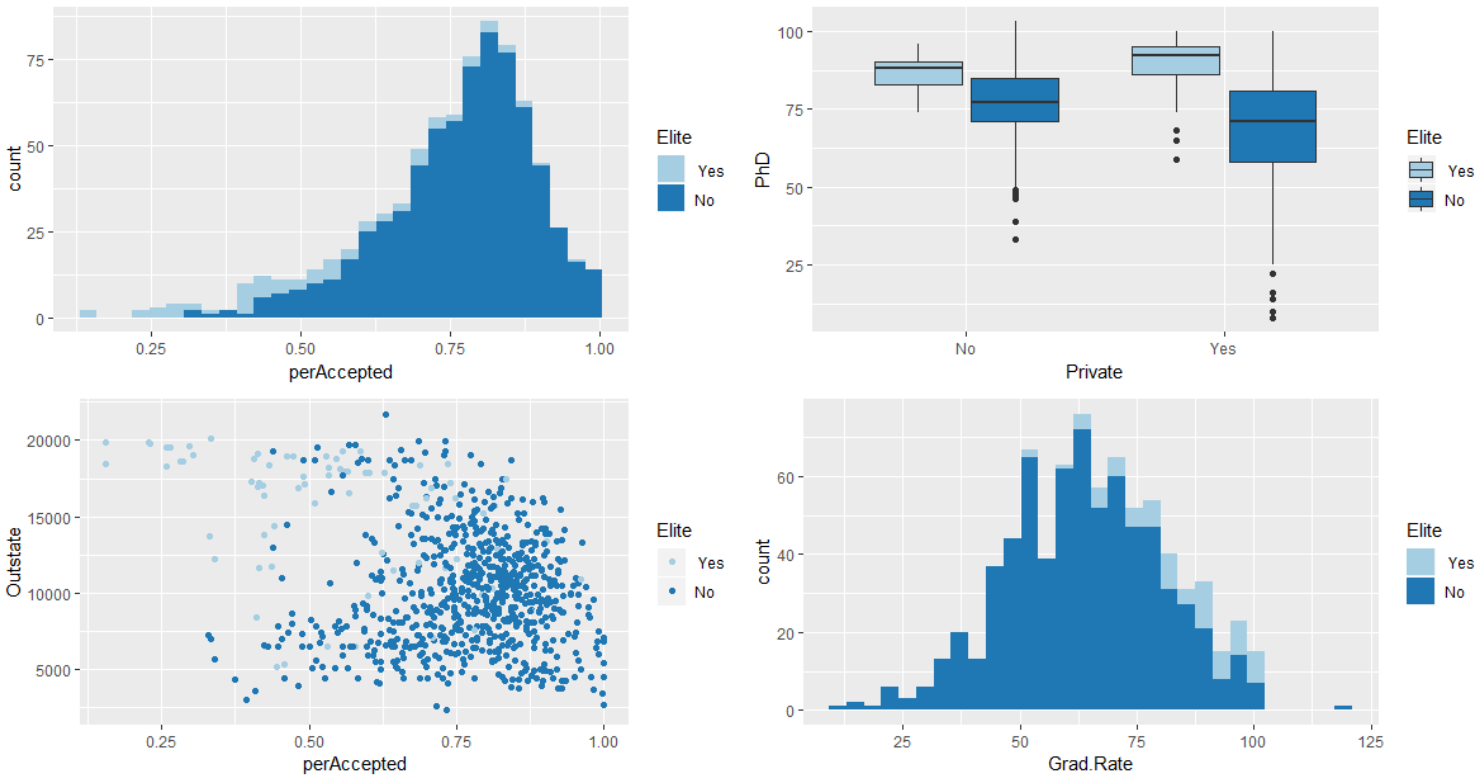


vi. I created a "perAccepted" variable that gives the percentage of applicants who were accepted out of the total number of applicants. From this, we can see that though only elite schools have very low acceptance rates (< 25%), there are still some schools with fairly large acceptance rates (>75%).

Further, we can look to see whether these elite schools vary by whether they are private or not, and whether they have a higher portion of faculty with PhDs. From this we can see that there is a big difference in terms of the variance between non-elite schools and elite schools. Non-elite schools have a larger range, with private non-elite schools having almost no faculty with PhDs to having 100% with PhDs. Non-elite schools do have a slightly lower median in the percentage of faculty PhDs, but there does not appear to be much of a difference across private and non-private schools.

Using the percent accepted variable again, we can see whether elite schools do in fact have smaller acceptance rates and their associated tuition costs. Elite schools who accept few students also tend to be the most expensive, though there are a few elite, lower-cost schools.

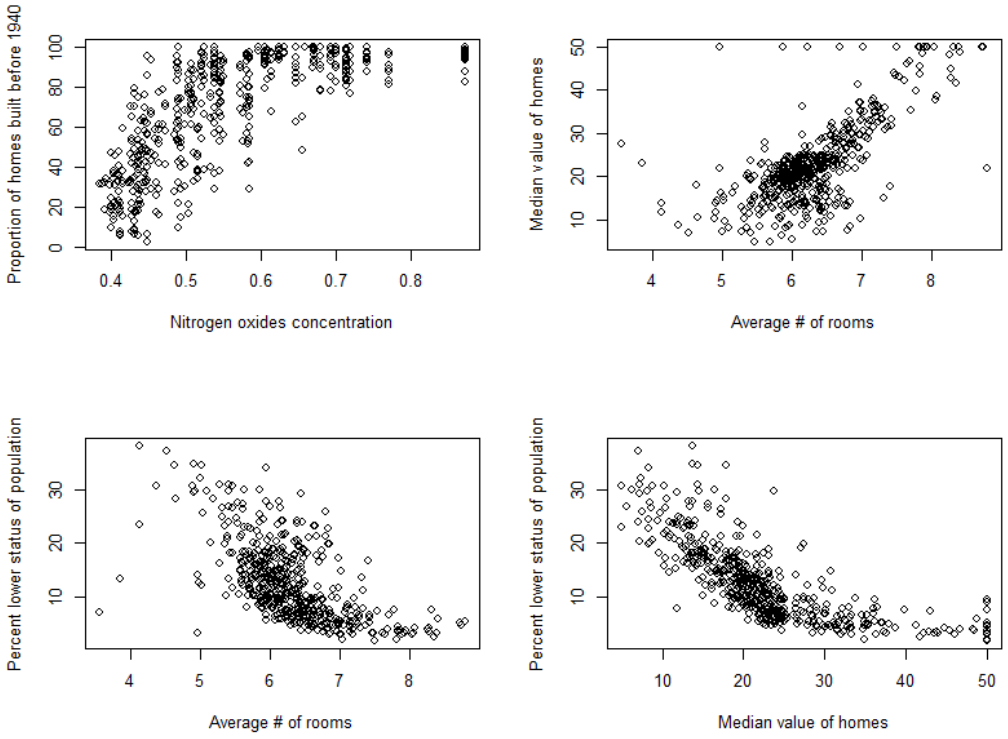
Lastly, we can determine whether the graduation rate of these elite schools is better than that of non-elite schools. The histogram does show that elite schools all have a graduation rate of 50% or greater.



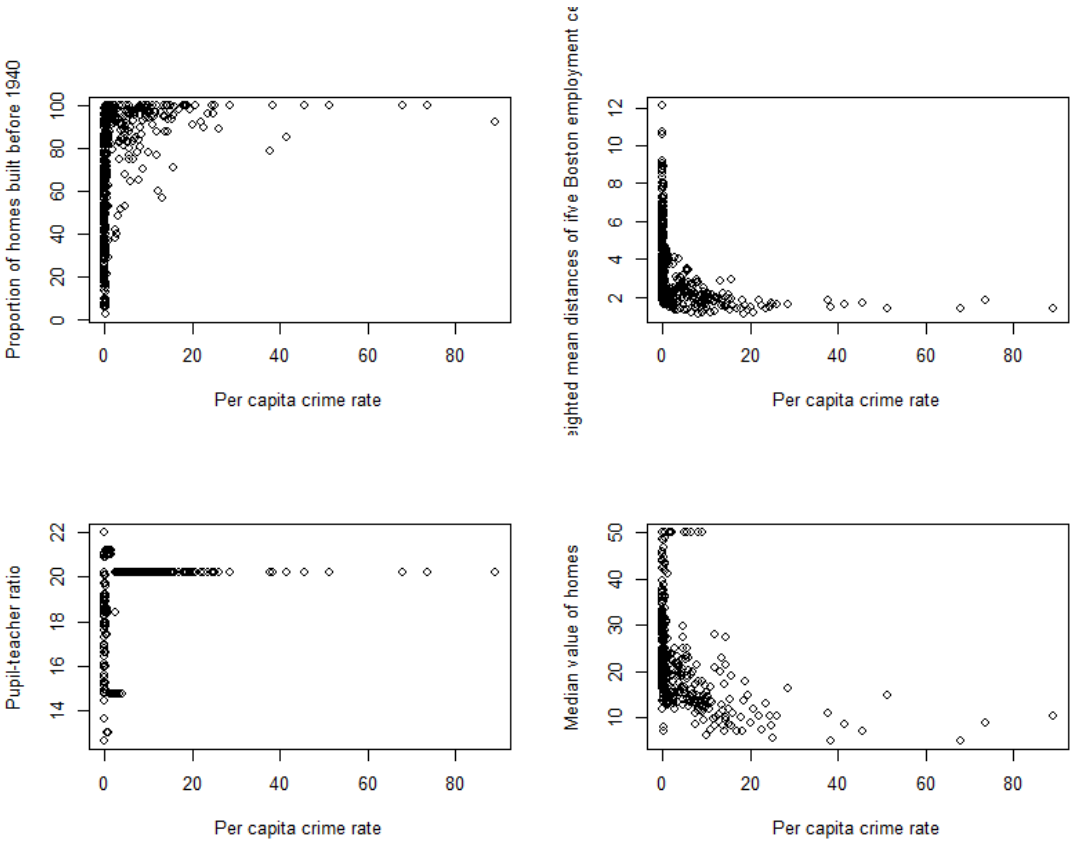
10.

a. This dataset has 506 rows (observations) and 14 columns (variables).

b. As nitrogen oxides concentration increases, so does the proportion of homes built before 1940. As the average number of rooms in a dwelling increases, it unsurprisingly also increases the median value of homes in that suburb. Further, the average number of rooms in a dwelling also appears to have a negative relationship with the percent of "lower status population" in the suburb. Similarly, as the median value of the home increases, the percent of "lower status of population" also decreases.



c. Though not a linear relationship, the per capita crime rate appears to have a positive association with proportion of homes built before 1940. Again, though not a linear relationship, the crime rate also increases when the weighted mean distance of Boston employment centers is smaller. Interestingly, all suburbs with a high crime rate had an average class size of 20 students. Lastly, though again not a linear relationship, the crime rate among suburbs with low median home values is higher.



d. 35 suburbs in this data set bound the Charles river.
e. The median pupil-teacher ratio of Boston suburbs is 19.05.
f. The suburb with observation number 399 has the . It is 7th highest in crime rate, has the maximum index of accessibility to radial highways, the lowest median value of owner-occupied homes, one of the towns with the largest proportion of blacks, and all owner-occupied homes were built before 1940.
g. 64 suburbs average more than 7 rooms per dwelling and 13 average more than 8 rooms. Suburbs with an average of more than 8 rooms per dwelling on average have a lower crime rate (0.72) compared to the overall (3.61). The tax rate is lower at 325.08 in this group compared to the overall tax rate of 408.24, and the percentage of "lower status of the population" is only 4.31 compared to the mean of 12.65 for all Boston suburbs. The mean of the median value of owner-occupied homes is also nearly double that of the overall at 44.2 compared to 22.5.