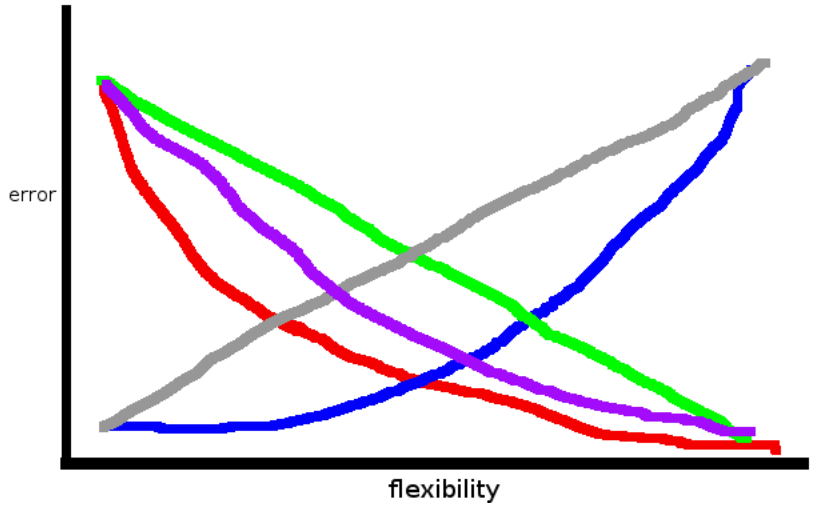
Math 650 HW 1

Mose Wintner

1.a) An inflexible method would perform better; a flexible method on a large sample size with few covariates will overfit the data.
b) Conversely, an flexible method on a small sample size with many covariates will underfit the data, so an inflexible method will perform better.
c) A flexible method is better; choosing an inflexible model like linear least squares or even the wrong parametric model will be insufficient to fit the data.
d) An inflexible method is better; one that is relatively flexible will fit the model to the errors, i.e. overfit, which we don't want.

2. a) Regression, inference. $n = 500$, $p = 4$.
b) Classification, prediction. $n = 20$, $p = 14$.
c) Regression, prediction. $n = 52$, $p = 4$.

3.



red: (squared) bias
blue: variance
green: training error
gray: test error
purple: Bayes (irreducible) error

4. a) i. Take images of single objects from the internet and determine the object. The response is likely a vector of scores corresponding to names of different objects. The predictors are likely outputs of a convolutional neural network whose input may just be a raw image file; features (predictors) are inferred from training data (i.e. image-label pairs). Inference.
ii. Classify origin of a dish (response) from ingredient list (and training data, i.e. many ingredient list-origin pairs) (predictors). Inference.

iii. Predict election outcome in a certain district (response) based on median income, number of schools/universities, age distribution, and outcomes of past elections (predictors). Prediction.

b) i. Attempt to predict a child's future income bracket (response) by parental income, school performance metrics, and hometown (predictors). Prediction.
ii. Calculate independent variables x containing prisoner data like crime severity, amount of extant family, money, home location, quantifiable methods of their incarceration and rehabilitation, etc. to predict $f(x)$, likelihood of recidivism. Could be prediction or inference.
iii. Calculate commercial inequality confidence intervals based on income statistics; i.e. a minority individual in community X, or at company X, is likely to earn a fraction $0 < f_X(y) \leq 1$ of that of a majority individual based on stratifying covariates y. ANOVA would probably be used here. Inference.

c) i. Inferring chemical composition of a star using spectrographic data.
ii. Inferring which states of medical covariates in a patient are signals of onset of an advanced stage of a disease.
iii. Inferring, based on a study of success/failure, which states of covariates were most likely to lead to success.

5. A very flexible approach for statistical learning is appropriate when we suspect a complex relationship between predictors and response and/or know very little about its form. The flexibility will allow for complexity in the function f we are approximating (discrete in the case of classification). An inflexible approach is appropriate when we suspect the relationship between predictor and response to be relatively simple, i.e. linear or parametric. It's also more appropriate when we suspect the data to contain many outliers or the error variance is large, because a flexible approach could overfit the model to the outliers/error. It's also necessary when we need the model to interpretable, i.e. when we want details of the relationship between the covariates and response. Generally, to avoid overfitting, we decrease the flexibility of our method. To avoid underfitting, we increase it.

6. A parametric approach is good when we want the function we're learning to be a member of, or well-approximated by, a member of a parametric family of functions, like $\{e^{ax} : a > 0\}$. A nonparametric approach is really a parametric approach with many parameters, for example, spline coefficients. If we expect the function we're training to be complex enough to need the freedom to possibly take the form of an arbitrary spline, then a nonparametric method is desirable.

7. a)

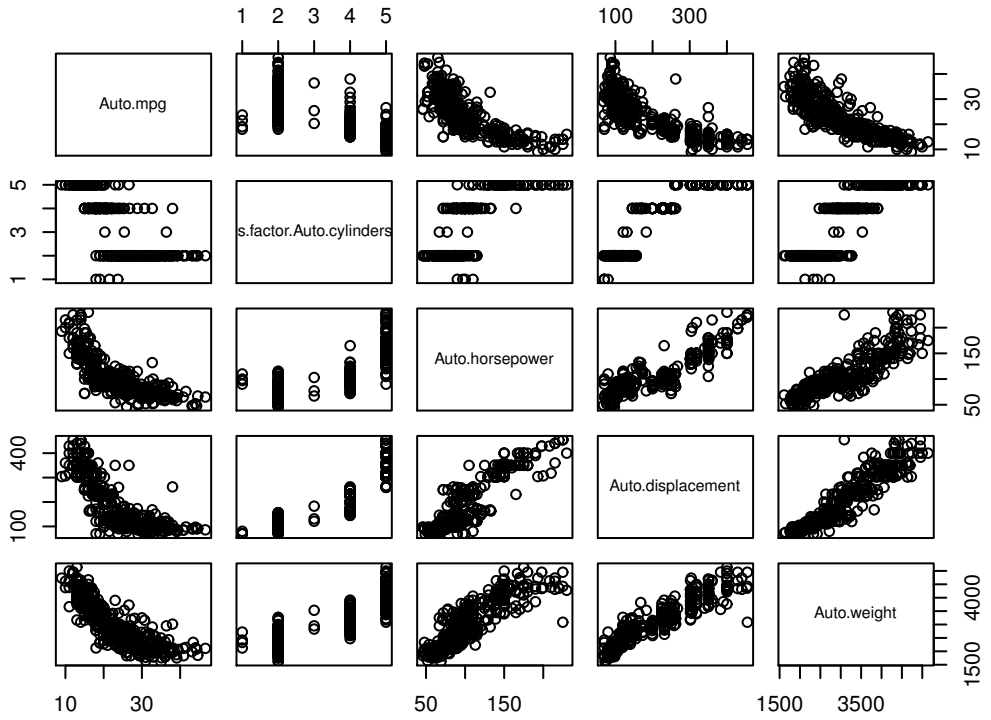
obs.	d
1	9
2	4
3	10
4	5
5	2
6	3

b) Green, because the test point is closest to observation 5.
c) Red. Of the 3 nearest neighbors, the majority are red.
d) Large, so the method is more flexible. A small K makes for a more rigid decision boundary.

8. *i*) through *v*) see attached .r file.
vi) Cazanovia College has a graduation rate $> 100\%$. The personal expenditure column has an

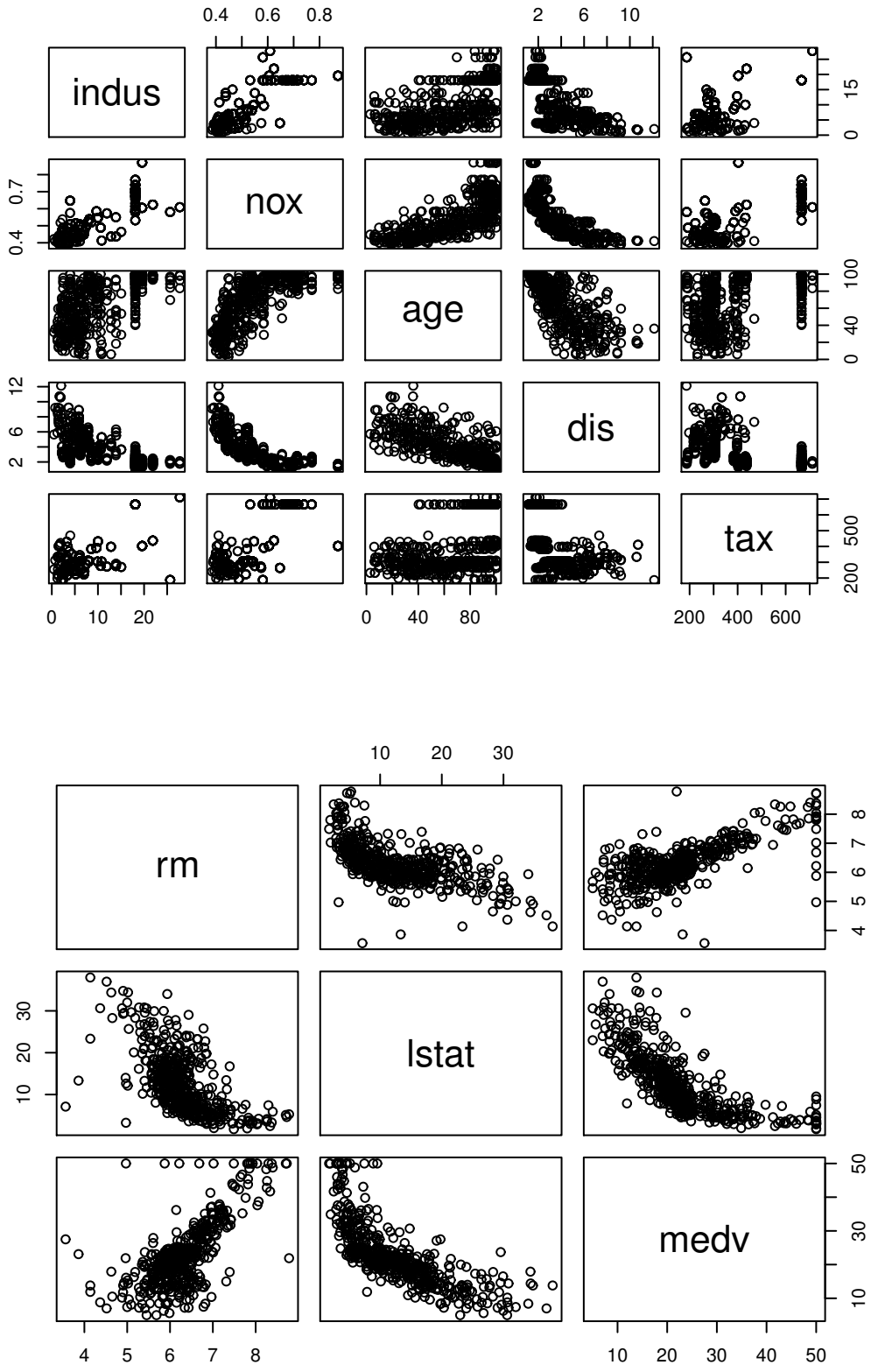
extremely high variance. After removing the very expensive schools from the pool, there appears to be no correlation between % of professors with terminal degree and instructional expenditure.

9. a) origin and name are the only non-quantitative variables.
b), c), d) see .r file.
e)



Weight is positively correlated with number of cylinders, horsepower, and displacement; it is negatively correlated with mpg, all of which are to be expected. The rest of the patterns above are predictable: mpg is negatively correlated with all other variables. And the other variables are all positively correlated with each other. Acceleration is also correlated with some other covariates but not all.
f) Yes. mpg is negatively correlated, and fairly strongly at that, with the other variables above.

10. a) 14 columns, 506 rows. The columns, covariates; the rows, towns near Boston.
b) Aided by cor(),



Industrial areas and neighborhoods with old homes are more likely to have more nitrogen oxide air pollution. Mean distance to Boston employment centers is negatively correlated with nitrogen oxide air pollution. There are also the expected correlations between number of rooms per dwelling, lower status of the population, and median value of occupied homes. Namely, the more poor people, the

more likely a town is to have occupied large homes, possibly because families live together. Also, the more poor people, the less valuable the homes are and vice versa.
c) From cor(), highway accessibility and property tax rate are the best (linear) predictors of crime rate. Good highway accessibility could also mean "inner city". And crime rate might well be a factor for property insurance actuaries.
d) There are 8 suburbs with a per capita crime rate above 40, which is very high. Three of those have per capita crime rates above above 65, which is almost impossibly high. The worst is 88 crimes per capita. The lowest crime rate is .00632 crimes per capita, so the range is enormous, with many outliers. 137 towns have tax rates of at least \$666 per \$10,000, and the lowest is \$187, so the range is not quite as dramatic but still very large. The pupil-teacher ratio ranges from 12.6 to 22.0, so not such a range, and very few outliers.
e) 35
f) 19.05
g) Suburb 399 has the lowest median owner-occupied home value. The suburb has a high crime rate, no large lots, a high number of average rooms per dwelling, every single home built prior to 1940, a short distance to Boston employment centers, a very high accessibility to radial highways (if I'm interpreting the "index" correctly), a very high property tax rate, a high parent-teacher ratio, a relatively high proportion of black residents, and a very high proportion of lower-class residents.
h) 64 towns average more than 7 rooms per dwelling. 13 average more than 8 rooms. Predictably, the median property value is high. Most of those homes appear to be old, built before 1940. This could be a result of a particular immigration pattern that happened sometime before 1940. The neighborhoods also appear to be mostly black neighborhoods. They are not predominantly lower-class, and have a low concentration of nitrogen oxides in the air. The rest of the covariates are pretty stratified.