2. ISLR: 2.4.1

(a) The sample size n is extremely large, and the number of predictors p is small.

Performance of flexible statistical learning method will be "better", as a flexible model will work better with a large sample size.

(b) The number of predictors p is extremely large, and number of observation n is small.

"worse": As a flexible model is very likely to overfit. It will learn detail and noise from data.

(c) The relationship between predictors and response is highly non-linear.

"better": As flexible models have higher degrees of freedom. It is likely to perform well on overly highly non-linear data.

(d) The Variance of the error terms; i.e $\sigma^2 = Var(\varepsilon)$, is extremely high.

"worse": Contrary to an inflexible model, a flexible model will fit the noise in the error terms. Thus, with a different sample, we will have a different fit. A simpler model will do bett a better estimation of f.

Q.3 ISLR 2.4.7

(a) Compute the Euclidean Distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$

X_1	X_2	X_3	Obs	Euclidean dist
0	3	0	1 R	$= \sqrt{0^2+3^2+0^2} = 3$
2	0	0	2 R	$= \sqrt{2^2} = 2$
0	1	3	3 R	$= \sqrt{1^2+3^2} = \sqrt{10}$
0	1	2	4 G	$= \sqrt{1^2+2^2} = \sqrt{5}$
-1	0	1	5 G	$= \sqrt{2}$
1	1	1	6 R	$= \sqrt{3}$

(b) What is our prediction with $K=1$? Why?

"Green": As point $(X_1 = -1, X_2 = 0, X_3 = 1)$ is the 1st closest neighbour to $(X_1 = 0, X_2 = 0, X_3 = 0)$

(c) What is our prediction with $K=3$? Why?

G R R

Closest 3 neighbours :- Obs 5, 6, 2

"Red": The probability of test point belonging to "Red" is higher than the probability of it belonging to "Green".

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

(less neighbours)

"small": Smaller values of K means a more flexible model. A flexible model will produce a more non-linear decision boundary. If K is large, more data points are considered, hence we get a more linear decision boundary.