

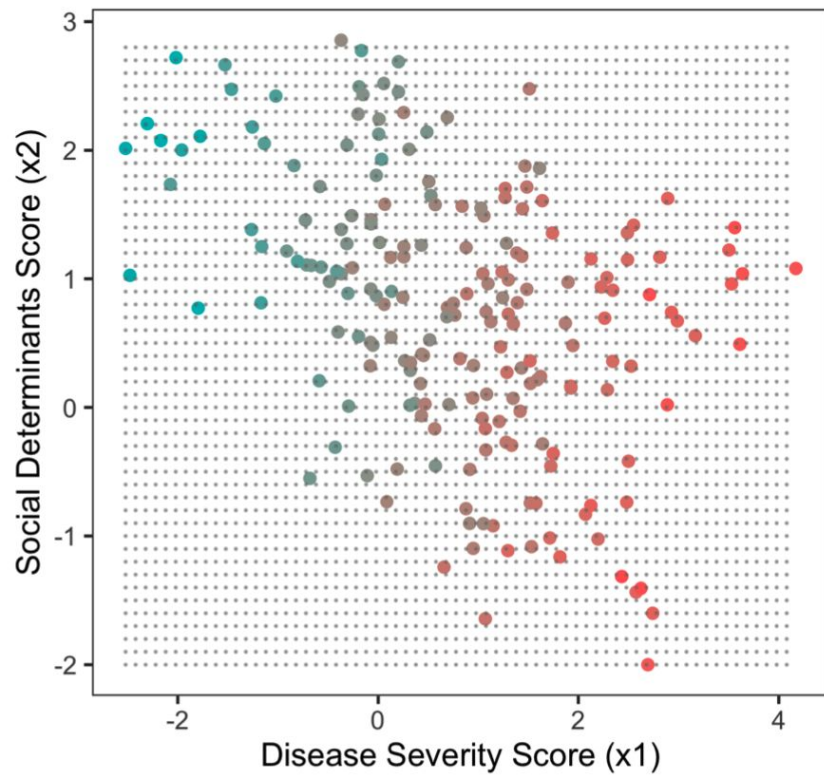
Chapter 3: The Basics of Regression

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor




How to Use this Guide

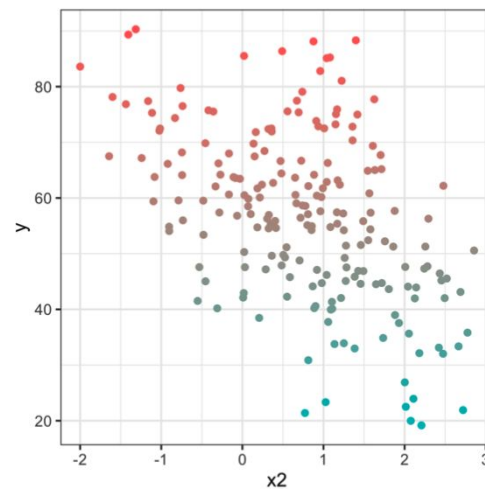
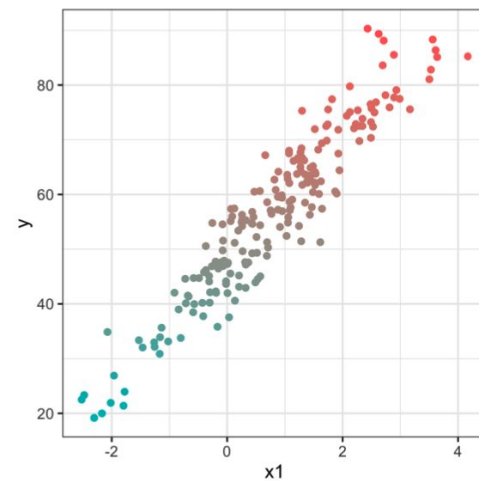
- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes



Recurrence
Biomarker (y)

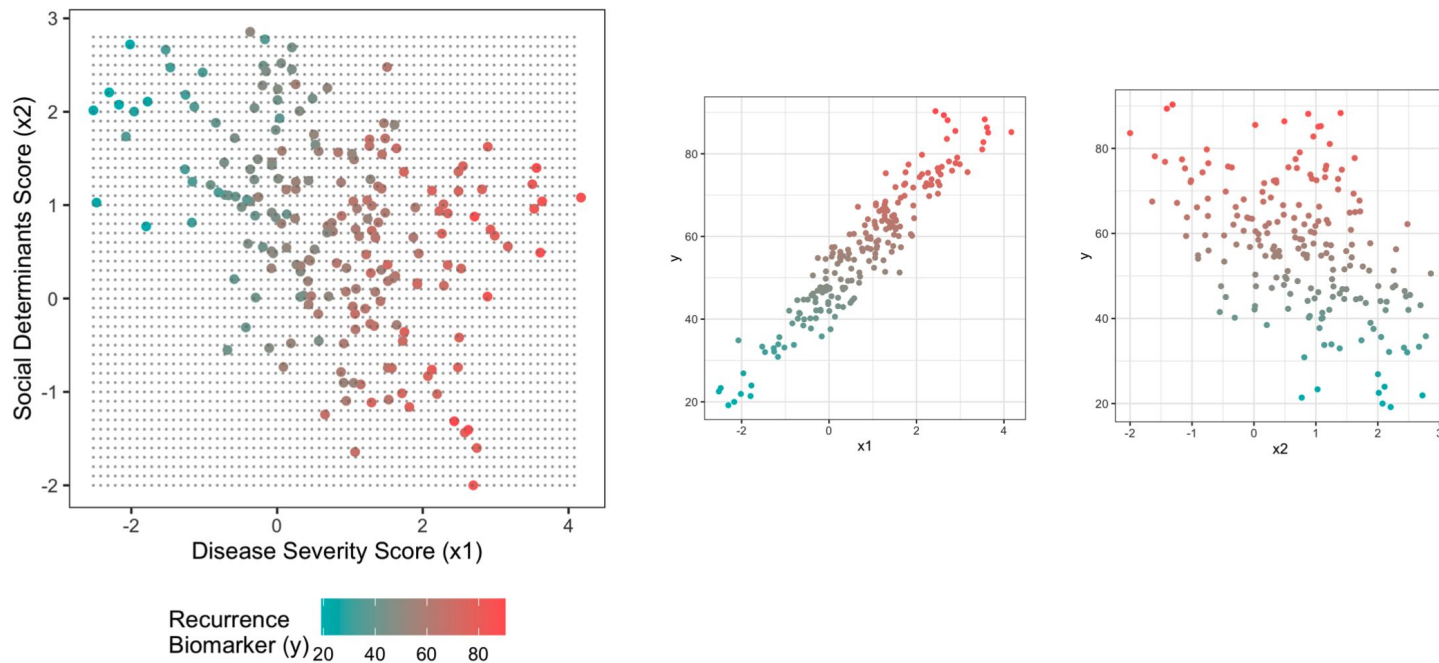


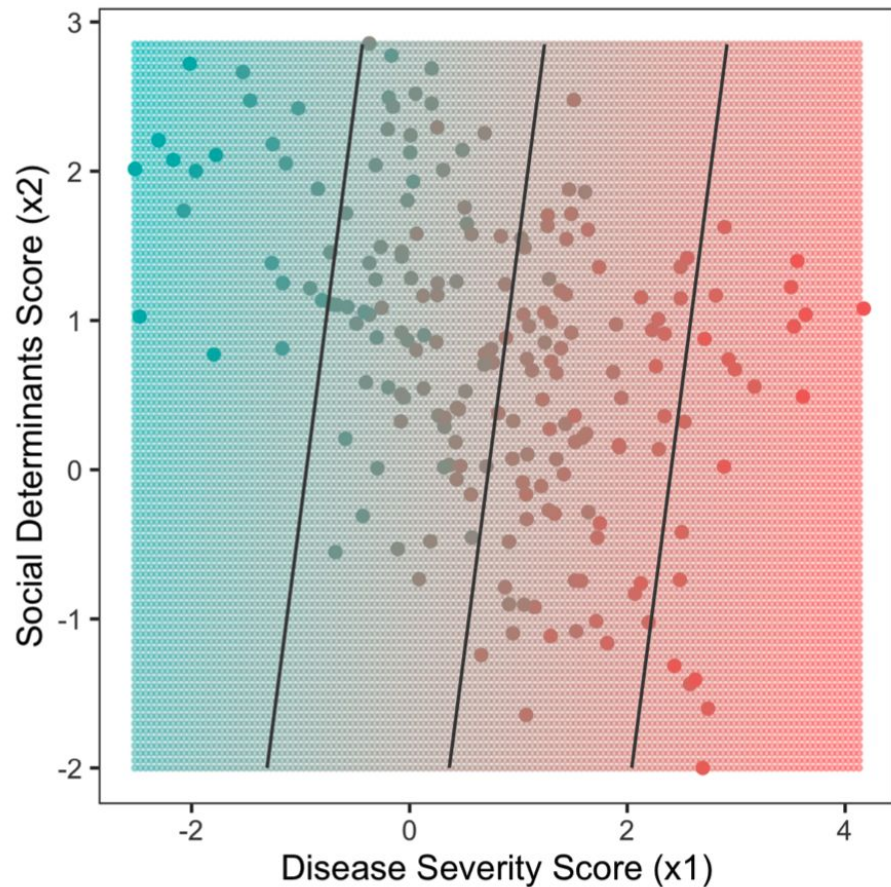
20 40 60 80



Question 3.1

Which of the two predictors, x_1 or x_2 , appears to more strongly influence the value of the recurrence biomarker? Explain your reasoning using evidence from the preceding three plots.





Call:

```
lm(formula = y ~ x1 + x2, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.9218	-3.1032	0.2891	2.8316	12.5813

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.8600	0.5370	92.844	< 2e-16 ***
x1	10.4372	0.2855	36.555	< 2e-16 ***
x2	-1.8824	0.3609	-5.215	4.63e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.769 on 197 degrees of freedom

Multiple R-squared: 0.9026, Adjusted R-squared: 0.9016

F-statistic: 912.4 on 2 and 197 DF, p-value: < 2.2e-16

$$\hat{y} = 49.8600 + 10.4372x_1 - 1.8824x_2$$

Question 3.2

Compare and contrast the output from the linear regression model with the output from the logistic regression model in Chapter 2. What looks the same? What looks different? What is being predicted in each case?

Linear Regression

Call:

```
lm(formula = y ~ x1 + x2, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.9218	-3.1032	0.2891	2.8316	12.5813

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.8600	0.5370	92.844	< 2e-16 ***
x1	10.4372	0.2855	36.555	< 2e-16 ***
x2	-1.8824	0.3609	-5.215	4.63e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.769 on 197 degrees of freedom

Multiple R-squared: 0.9026, Adjusted R-squared: 0.9016

F-statistic: 912.4 on 2 and 197 DF, p-value: < 2.2e-16

Logistic Regression

Call:

```
glm(formula = y ~ x1 + x2, family = "binomial", data = df)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.88232	-0.90614	-0.05965	0.86579	2.28489

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.9780	0.2945	3.321	0.000897 ***
x1	0.1344	0.1372	0.980	0.327272
x2	-1.3981	0.2316	-6.035	1.59e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

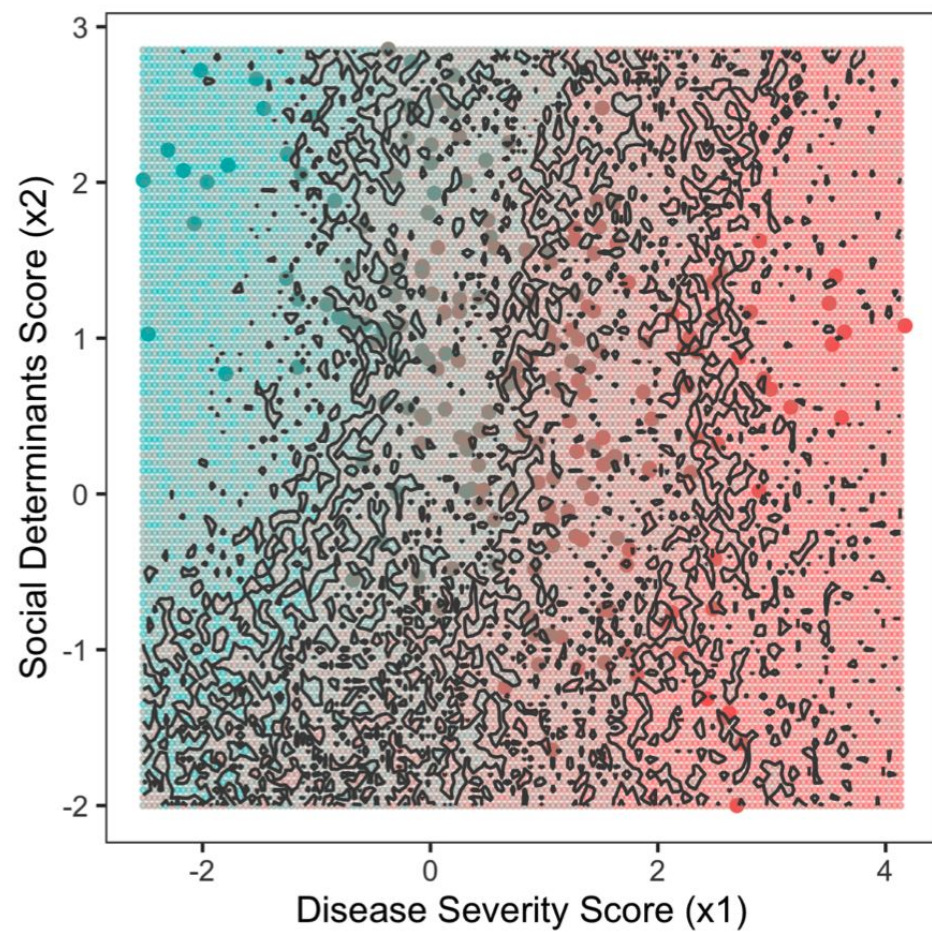
(Dispersion parameter for binomial family taken to be 1)

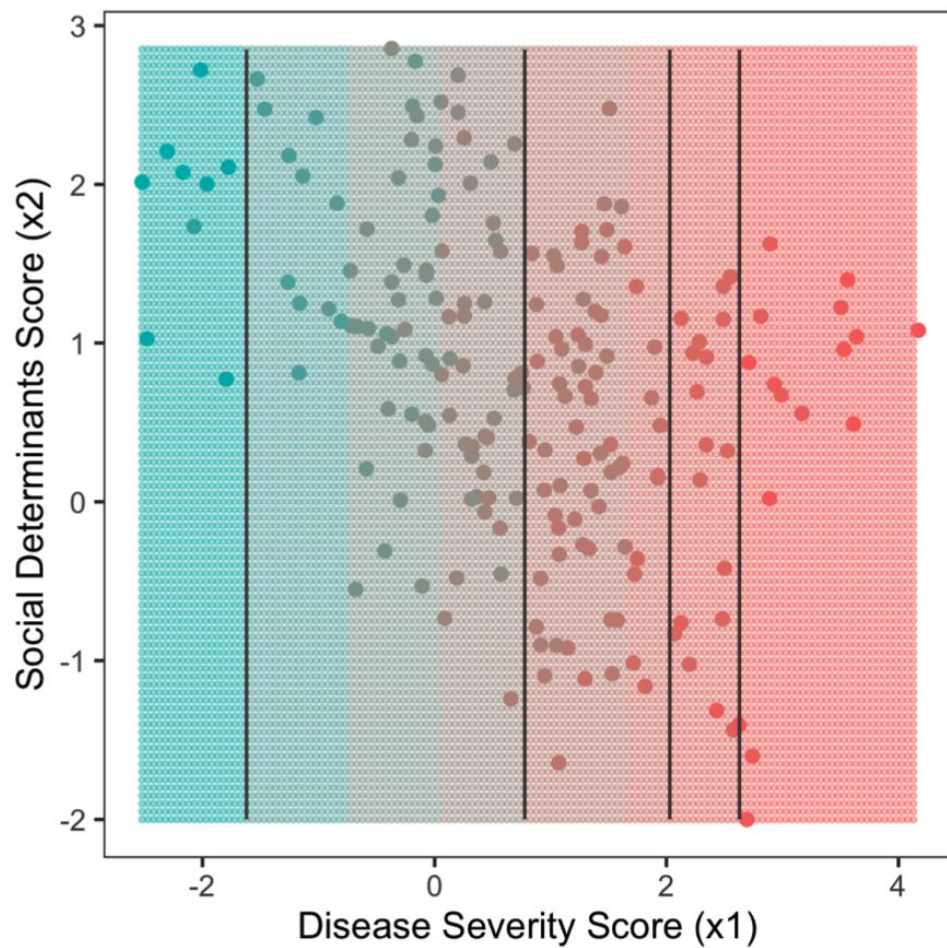
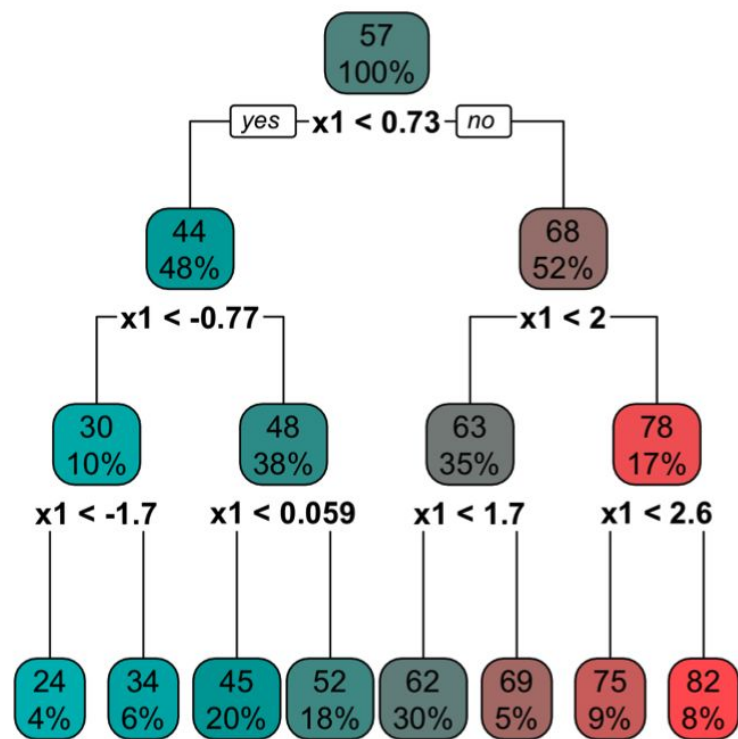
Null deviance: 277.26 on 199 degrees of freedom

Residual deviance: 209.54 on 197 degrees of freedom

AIC: 215.54

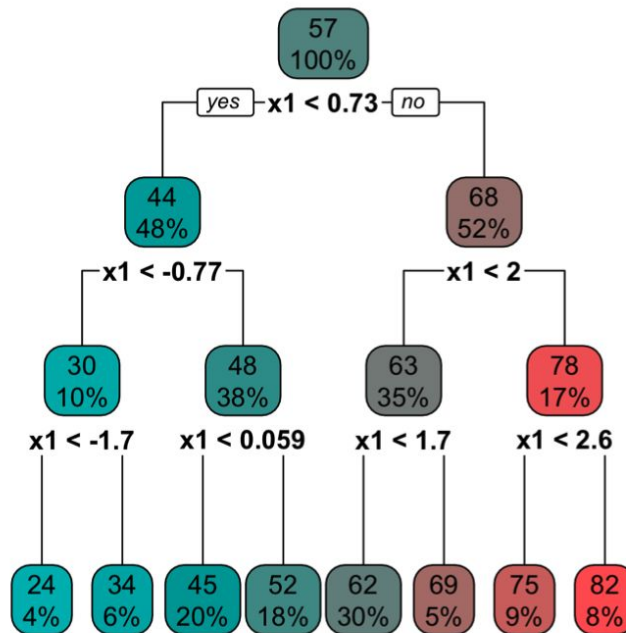
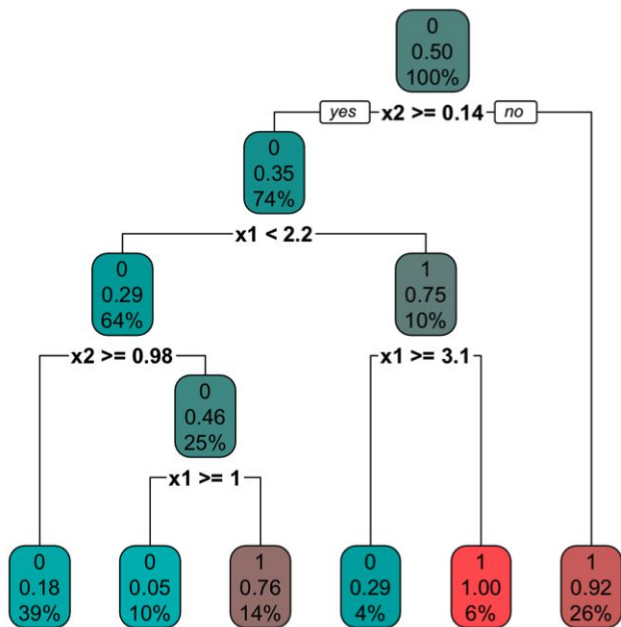
Number of Fisher Scoring iterations: 4





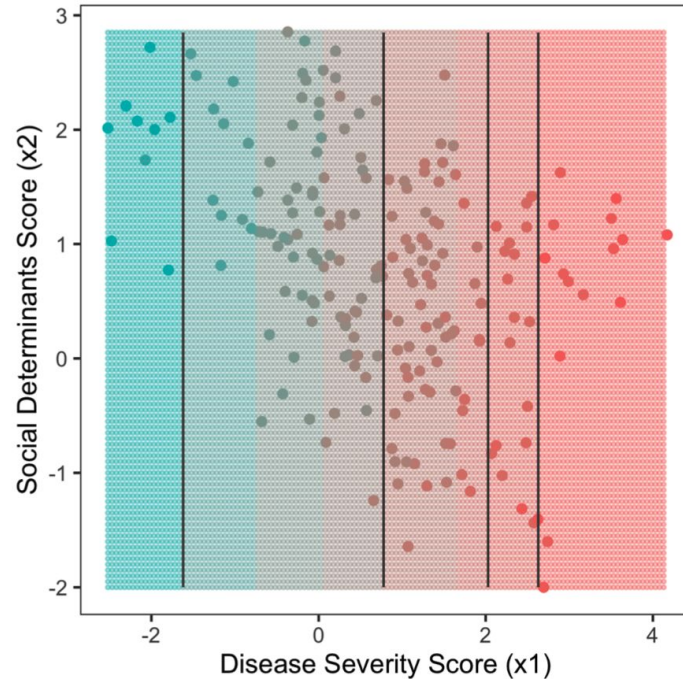
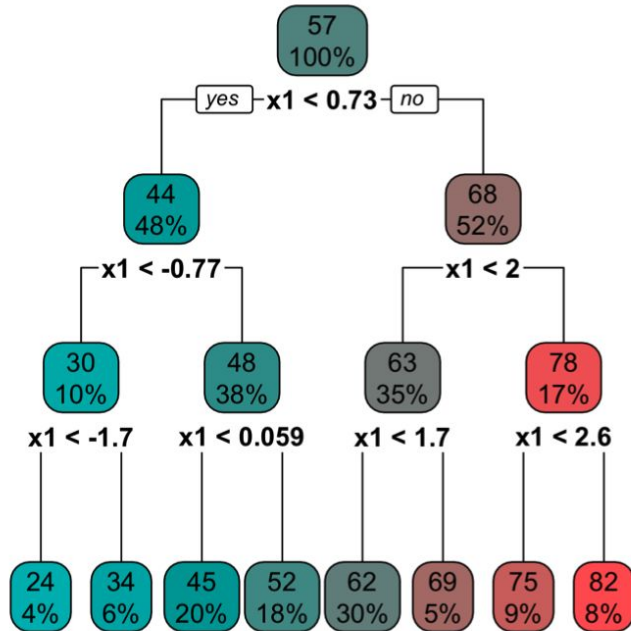
Question 3.3

Compare this decision tree with the decision tree for the classification problem in Chapter 2. What is the same? What is different?



Question 3.4

This **regression tree** has eight leaves. What region of the feature space does each leaf correspond to?



Question 3.5

What are the advantages and disadvantages of each of these three regression algorithms (linear regression, KNN, regression tree)?

