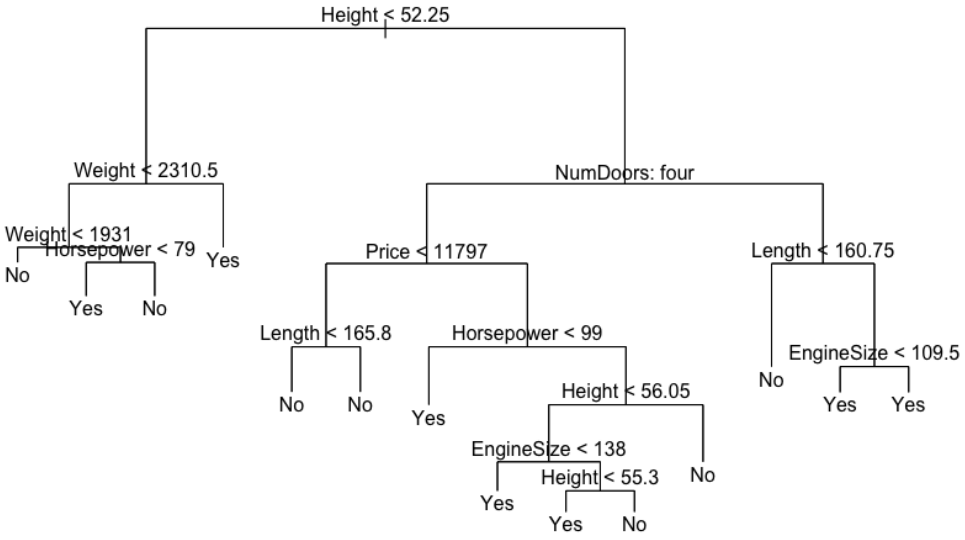
Data Analytics Assessment #4
Beiming Zhang

Programming Part

Q1 a)
Output:

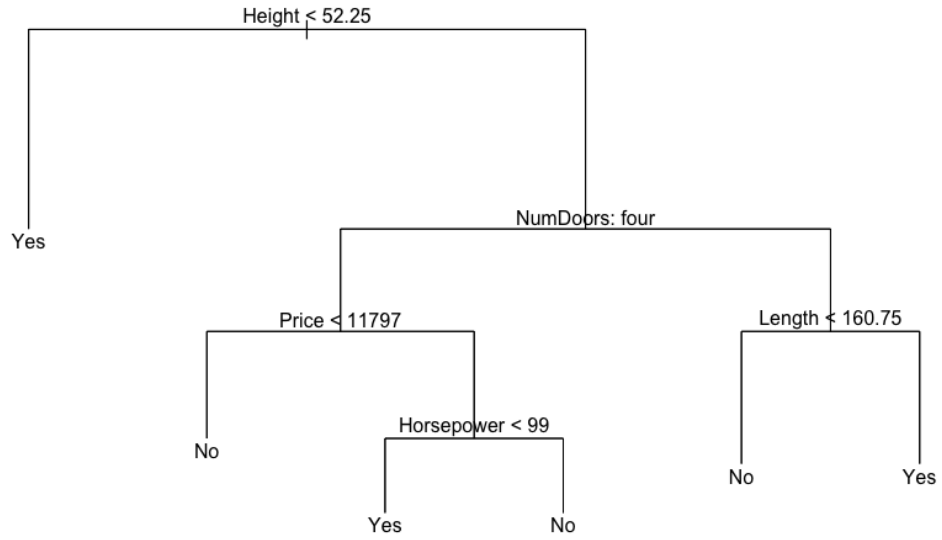


- The decision tree is as above.
- The predictors used at the nodes of the tree are **“Height”, “Weight”, “NumDoors”, “Horsepower”, “Price”, “Length” and “EngineSize”**. The tree has **14 terminal nodes (leaves)**.

Q1 b)
Output:

```
$size
[1] 14 11  8  6  4  2  1

$dev
[1] 32 32 33 32 34 50 75

$k
[1] -Inf  0.0  1.0  2.5  3.0  8.5 28.0

$method
[1] "misclass"

attr(,"class")
```

[1]"prune" "tree.sequence"



- The deviance is smallest when size is 14, 11, and 6, which are all 6. According to Occam's razor principle, the **best tree size is 6**.
- The decision tree after pruning is as above.

Q1 c)

- Height=60, the condition Height<52.25 is false; NumDoors: Two, the condition NumDoors: Four is false; Length=160, the condition Length<160.75 is true, the decision tree returns "No".
- According to the best tree, his car will **not** incur a high loss.

Q2 a)

Output:

	titanic_test_survived	
knn_k2	0	1
0	0.55605381	0.13901345
1	0.06278027	0.24215247

	titanic_test_survived	
knn_k4	0	1
0	0.55605381	0.14349776
1	0.06278027	0.23766816

	titanic_test_survived	
knn_k6	0	1

0	0.56950673	0.14573991
1	0.04932735	0.23542601

- 0.13901345+0.24215247=0.3812
- 0.06278027+0.24215247=0.3049
- 0.14349776+0.23766816=0.3812
- 0.06278027+0.23766816=0.3004
- 0.14573991+0.23542601=0.3812
- 0.04932735+0.23542601=0.2848
- When **K=2**, the survival rate in the test data is **38.12%** and the predicted survival rate is about 30.49%; when **K=4**, the survival rate in the test data is **38.12%** and the predicted survival rate is about 30.04%; when **K=6**, the survival rate in the test data is **38.12%** and the predicted survival rate is about 28.48%.

Q2 b)

Output:

k=2	k=4	k=6
0.7982063	0.7937220	0.8049327

- When **K=2**, the accuracy rate is about **79.82%**; when **K=4**, the accuracy rate is about **79.37%**; when **K=6**, the accuracy rate is about **80.49%**。
- Therefore, when **K=6, it works best**.

Conceptual Part

Q1 a)

$$d = \sqrt{(x_1 - x_{i1})^2 + (x_2 - x_{i2})^2 + (x_3 - x_{i3})^2}$$

- Calculate the Euclidean distance as follows:

Obs.	1	2	3	4	5	6
Distance	6.16	5.92	5.20	**4.00**	**2.83**	**1.41**

- Since K=3, the decision is made here based on the three points **4, 5, and 6** with the closest Euclidean distance.

Q1 b)

- The most recent qualitative response of the three points 4, 5 and 6 are Medium, Medium and Low respectively.
- According to the principle of KNN, we choose **Medium**, which accounts for the majority, as the prediction of classification.

Q1 c)

- The most recent quantitative response of the three points 4, 5 and 6 are 110, 120 and 70 respectively.
- According to the principle of KNN, the average of Y values for the 3 nearest neighbors is **100**, which is our regression prediction.

Q2 a)

- For region R1, we have the following observations falling in this region:

Obs.	Y
3	Low
8	Low
10	Low
15	High

- **Region R1 is labeled as "Low".**
- For region R2, we have the following observations falling in this region:

Obs.	Y
1	High
2	Low
6	High
12	High

- **Region R2 is labeled as "High".**
- For region R3, we have the following observations falling in this region:

Obs.	Y
4	Low
7	High
11	Low

- **Region R3 is labeled as "Low".**

- For region R4, we have the following observations falling in this region:

Obs.	Y
5	High
13	Low
14	High

- **Region R4 is labeled as "High".**
- For region R5, we have the following observations falling in this region:

Obs.	Y
9	Low

- **Region R5 is labeled as "Low".**

Q2 b)

Obs.	Actual	Predicted	Classification
1	High	High	TP
2	Low	High	FP
3	Low	Low	TN
4	Low	Low	TN
5	High	High	TP
6	High	High	TP
7	High	Low	FN
8	Low	Low	TN
9	Low	Low	TN
10	Low	Low	TN
11	Low	Low	TN
12	High	High	TP
13	Low	High	FP
14	High	High	TP
15	High	Low	FN

- The confusion for the training dataset matrix is below:

	Predicted: "High"	Predicted: "Low"
Actual: "High"	5	2
Actual: "Low"	2	6

- The accuracy for the training dataset is (5+6)/15=**73.33%**

Q2 c)

Obs.	Actual	Region	Predicted	Classification
1	High	R3	Low	FN
2	Low	R4	High	FP
3	High	R2	High	TP
4	Low	R5	Low	TN
5	Low	R1	Low	TN

- The confusion for the training dataset matrix is below:

	Predicted: "High"	Predicted: "Low"

Actual: "High"	1	1
Actual: "Low"	1	2

- The accuracy for the training dataset is (1+2)/5=**60.00%**

Q3 a)

- For region R1, we have the following observations falling in this region:

Obs.	Y
3	1
8	9
10	1
15	13

- **The average of the Y values within R1 is 6.**
- For region R2, we have the following observations falling in this region:

Obs.	Y
1	19
2	6
6	17
12	18

- **The average of the Y values within R2 is 15.**
- For region R3, we have the following observations falling in this region:

Obs.	Y
4	5
7	13
11	6

- **The average of the Y values within R3 is 8.**
- For region R4, we have the following observations falling in this region:

Obs.	Y
5	19
13	6
14	11

- **The average of the Y values within R4 is 12.**
- For region R5, we have the following observations falling in this region:

Obs.	Y
9	8

- **The average of the Y values within R5 is 8.**

Q2 b)

Obs.	Y	Region	$\hat{Y}$	$(Y- \hat{Y})^2$
1	11	R3	8	9
2	9	R4	12	9
3	15	R2	15	0
4	7	R5	8	1

5	7	R1	6	1
Total	49		50	20

- For 5 observations:
- **RSS=20**
- **MSE=20/5=4**