BUAN 6356.002
Problem Set 5
Huyen Nguyen (Htn180001)

Question 1:
1. Data generation process:
- Create 500 records for each group
- Set seed = 75080
- Create 500 random values for z
- Create 500 random values for w
- Create 500 values for income (x) of group 1: x= 5*z + 50
- Create 500 values for SAT (y) of group 1: y= -100*z + 1100 + 50*w
- Round up value of Y
- Set value of Y if less than 200 to be 200 - the min of SAT
- Set value of Y if > 1600 to be 1600 - the max of SAT
- Create data for group 1 [id: 1, 500]

- Repeat same process to create data for group 2 [id: 501, 1000] with x= 5*z + 80 and y= -80*z+ 1200 + 50*w

- Repeat same process to create data for group 3 [id: 501, 1000] with x= 5*z + 30 and y = -120*z+ 1000 + 50*w

- Merge 3 data sets together

2.
Pooled OLS model:
Call:
lm(formula = sat ~ income, data = dtable)

Residuals:
```
    Min      1Q  Median      3Q     Max
-452.84  -81.64    7.67   88.71  440.50
```

Coefficients:
```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  950.8914   9.1279     104.17   <2e-16 ***
income       2.7923    0.1593     17.53    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 129.1 on 1498 degrees of freedom
Multiple R-squared: 0.1703, Adjusted R-squared: 0.1697
F-statistic: 307.4 on 1 and 1498 DF, p-value: < 2.2e-16

Fixed-effects model:
Call:
lm(formula = sat ~ income + group - 1, data = dtable)

Residuals:
```
     Min       1Q       Median       3Q      Max
-165.106  -34.157   -0.242   34.979  189.967
```

Coefficients:

```
            Estimate      Std. Error   t value   Pr(>|t|)
income  -20.173       0.268        -75.26   <2e-16 ***
group1  2111.255     13.559       155.71   <2e-16 ***
group2  2812.183     21.518       130.69   <2e-16 ***
group3  1605.304     8.469         189.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.54 on 1496 degrees of freedom
Multiple R-squared:  0.9978,  Adjusted R-squared:  0.9978
F-statistic: 1.732e+05 on 4 and 1496 DF,  p-value: < 2.2e-16
```

Group 1:

```
Call:
lm(formula = sat ~ income, data = dtable[group == 1])

Residuals:
    Min      1Q  Median      3Q     Max
-165.747  -32.762   -0.778   36.061  156.371

Coefficients:
                Estimate      Std. Error    t value      Pr(>|t|)
(Intercept)   2095.3391    21.8084      96.08       <2e-16 ***
income          -19.8534       0.4352       -45.62      <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.39 on 498 degrees of freedom
Multiple R-squared:  0.8069,  Adjusted R-squared:  0.8065
F-statistic:  2081 on 1 and 498 DF,  p-value: < 2.2e-16
```

Group 2:

```
Call:
lm(formula = sat ~ income, data = dtable[group == 2])

Residuals:
    Min      1Q  Median      3Q     Max
-167.216  -32.930    0.719   33.225  193.402

Coefficients:
                Estimate      Std. Error    t value    Pr(>|t|)
(Intercept)   2505.1204    36.8883      67.91     <2e-16 ***
income          -16.3257       0.4613       -35.39    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.52 on 498 degrees of freedom
Multiple R-squared:  0.7155,  Adjusted R-squared:  0.7149
F-statistic:  1252 on 1 and 498 DF,  p-value: < 2.2e-16
```

Group 3:

```
Call:
lm(formula = sat ~ income, data = dtable[group == 3])
```

Residuals:

Min	1Q	Median	3Q	Max
-145.298	-33.165	0.884	33.176	138.831

Coefficients:

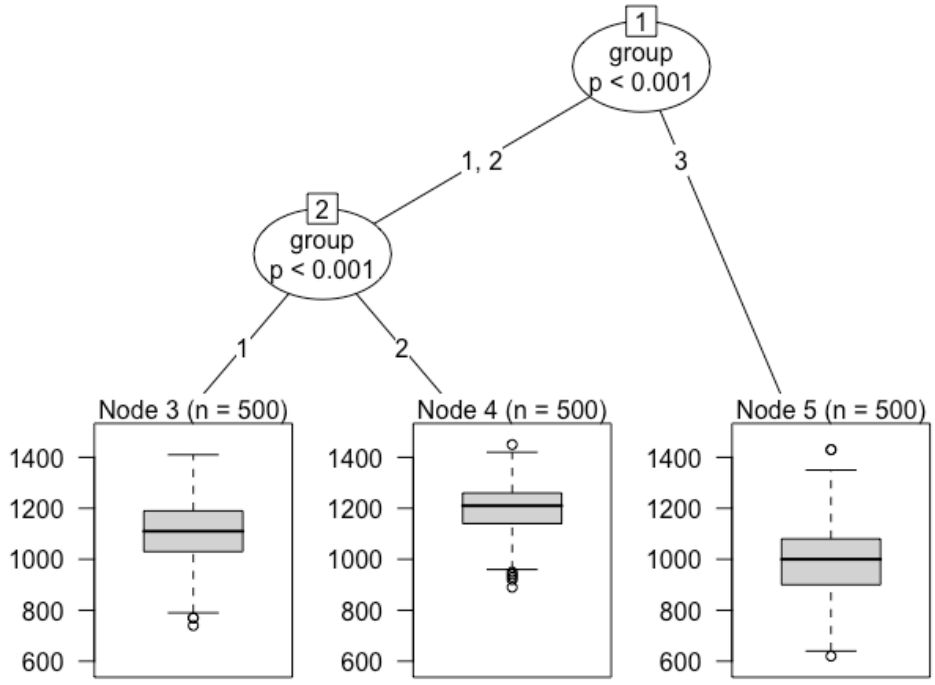
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1722.484	13.373	128.81	<2e-16 ***
income	-24.027	0.434	-55.36	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

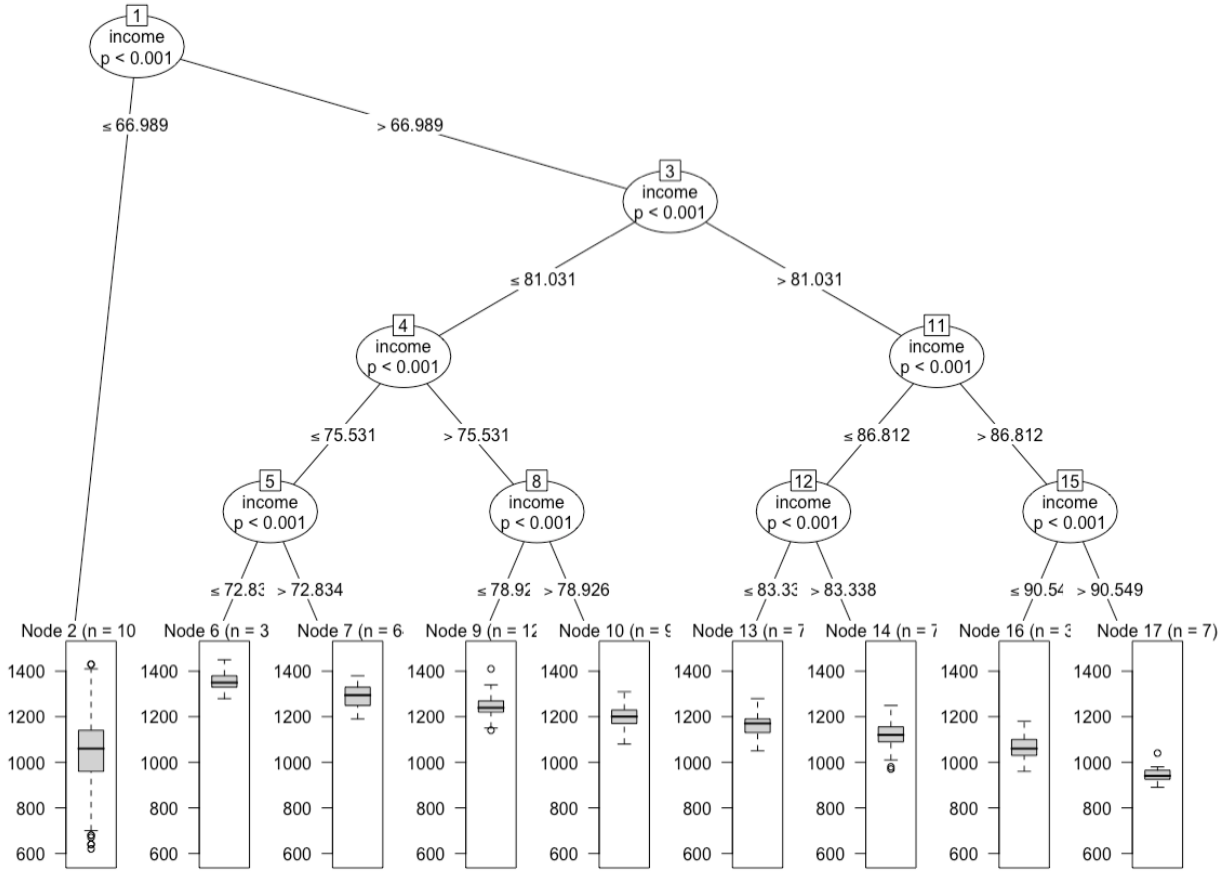
Residual standard error: 48.66 on 498 degrees of freedom
Multiple R-squared: 0.8602, Adjusted R-squared: 0.86
F-statistic: 3065 on 1 and 498 DF, p-value: < 2.2e-16

The signs are different between these models because the Pooled OLS model only considers income's effect to sat score of all 3 groups, while the Fixed-effects models considers income and effects of these 3 groups to the sat score. Each group's model will have different signs that reflect the effect of income to sat score of that group only.

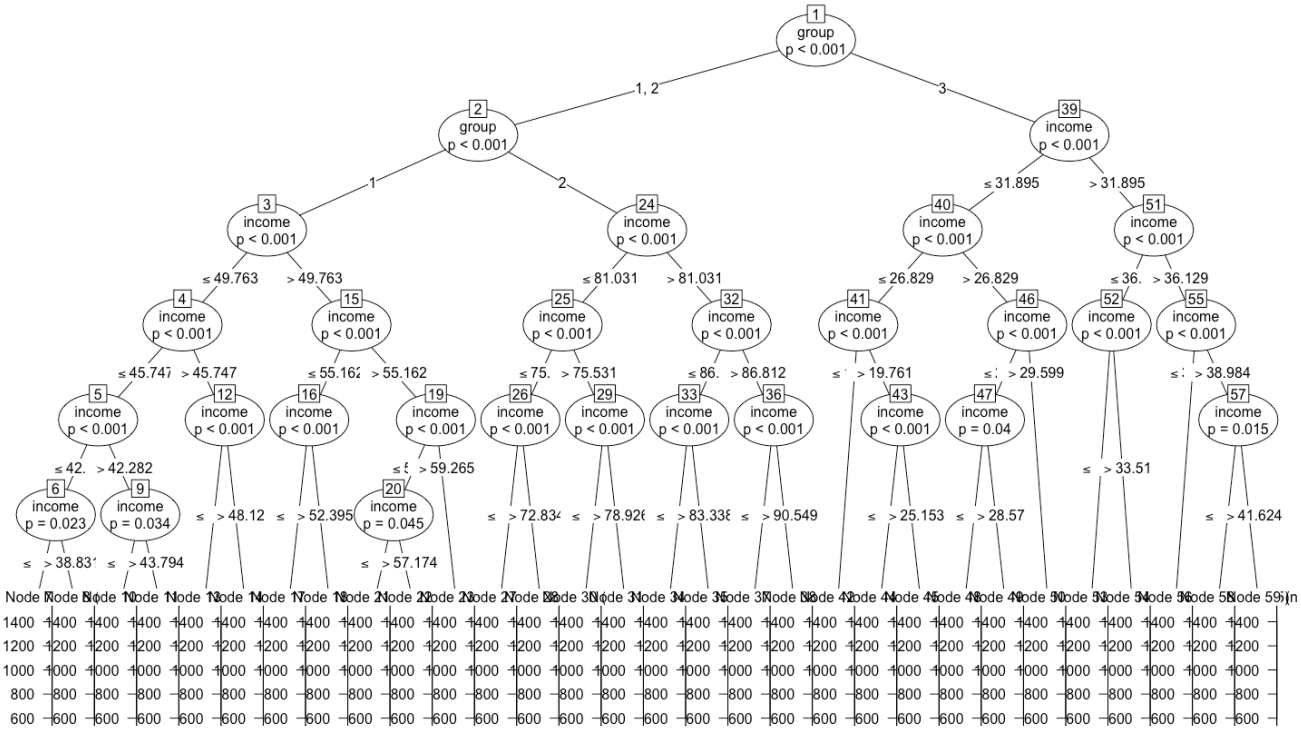
3.
Model SAT using group



Model SAT using income



Model SAT using both variables: income and group



We can see that model SAT using group generates 3 clusters according to 3 groups we created at the beginning. Model using income creates 9 clusters and model using both variables generates 31

clusters (divided by group first then by income). The more number of factors we use and the more diverse these factors, the more clusters we'll have.

4.
Generalized linear model tree (family: gaussian)

Model formula:
sat ~ 1 | income + group

Fitted party:
```
[1] root
|   [2] group in 2
|   |   [3] income <= 78.92595
|   |   |   [4] income <= 73.90014
|   |   |   |   [5] income <= 72.76712: n = 31
|   |   |   |       (Intercept)
|   |   |   |          1355.806
|   |   |   |   [6] income > 72.76712
|   |   |   |   |   [7] income <= 73.47808: n = 13
|   |   |   |   |       (Intercept)
|   |   |   |   |          1288.462
|   |   |   |   |   [8] income > 73.47808: n = 10
|   |   |   |   |       (Intercept)
|   |   |   |   |              1338
|   |   |   [9] income > 73.90014
|   |   |   |   [10] income <= 75.53141: n = 42
|   |   |   |       (Intercept)
|   |   |   |          1278.571
|   |   |   |   [11] income > 75.53141: n = 121
|   |   |   |       (Intercept)
|   |   |   |          1244.215
|   |   [12] income > 78.92595
|   |   |   [13] income <= 83.33764
|   |   |   |   [14] income <= 81.03089: n = 93
|   |   |   |       (Intercept)
|   |   |   |           1200.43
|   |   |   |   [15] income > 81.03089: n = 77
|   |   |   |       (Intercept)
|   |   |   |          1161.429
|   |   |   [16] income > 83.33764
|   |   |   |   [17] income <= 87.82086: n = 88
|   |   |   |       (Intercept)
|   |   |   |          1114.318
|   |   |   |   [18] income > 87.82086
|   |   |   |   |   [19] income <= 88.83985: n = 12
|   |   |   |   |       (Intercept)
|   |   |   |   |          1056.667
|   |   |   |   |   [20] income > 88.83985: n = 13
|   |   |   |   |       (Intercept)
|   |   |   |   |          983.8462
|   [21] group in 1, 3
|   |   [22] group in 1
|   |   |   [23] income <= 49.76334
|   |   |   |   [24] income <= 45.7473
```

```
|  |  |  |  |  [25] income <= 43.47638: n = 43
|  |  |  |  |      (Intercept)
|  |  |  |  |          1276.279
|  |  |  |  |  [26] income > 43.47638: n = 65
|  |  |  |  |      (Intercept)
|  |  |  |  |          1208.769
|  |  |  |  [27] income > 45.7473
|  |  |  |  |  [28] income <= 48.12029: n = 85
|  |  |  |  |      (Intercept)
|  |  |  |  |          1168.471
|  |  |  |  |  [29] income > 48.12029: n = 71
|  |  |  |  |      (Intercept)
|  |  |  |  |          1122.676
|  |  |  [30] income > 49.76334
|  |  |  |  [31] income <= 55.16208
|  |  |  |  |  [32] income <= 52.39517: n = 94
|  |  |  |  |      (Intercept)
|  |  |  |  |          1076.915
|  |  |  |  |  [33] income > 52.39517: n = 59
|  |  |  |  |      (Intercept)
|  |  |  |  |          1027.119
|  |  |  |  [34] income > 55.16208
|  |  |  |  |  [35] income <= 59.26473: n = 65
|  |  |  |  |      (Intercept)
|  |  |  |  |          972.9231
|  |  |  |  |  [36] income > 59.26473: n = 18
|  |  |  |  |      (Intercept)
|  |  |  |  |          847.7778
|  |  [37] group in 3
|  |  |  [38] income <= 31.89534
|  |  |  |  [39] income <= 26.82913
|  |  |  |  |  [40] income <= 19.76067: n = 12
|  |  |  |  |      (Intercept)
|  |  |  |  |          1308.333
|  |  |  |  |  [41] income > 19.76067
|  |  |  |  |  |  [42] income <= 25.03478: n = 53
|  |  |  |  |  |      (Intercept)
|  |  |  |  |  |          1158.302
|  |  |  |  |  |  [43] income > 25.03478: n = 52
|  |  |  |  |  |      (Intercept)
|  |  |  |  |  |          1098.846
|  |  |  |  [44] income > 26.82913
|  |  |  |  |  [45] income <= 29.59926: n = 105
|  |  |  |  |      (Intercept)
|  |  |  |  |          1043.048
|  |  |  |  |  [46] income > 29.59926: n = 92
|  |  |  |  |      (Intercept)
|  |  |  |  |          991.1957
|  |  |  [47] income > 31.89534
|  |  |  |  [48] income <= 36.12898
|  |  |  |  |  [49] income <= 33.79417: n = 68
|  |  |  |  |      (Intercept)
|  |  |  |  |          928.9706
|  |  |  |  |  [50] income > 33.79417: n = 60
|  |  |  |  |      (Intercept)
```

```
|   |   |   |   |         875.8333
|   |   |   |   [51] income > 36.12898
|   |   |   |   |   [52] income <= 38.9841: n = 36
|   |   |   |   |       (Intercept)
|   |   |   |   |         816.6667
|   |   |   |   |   [53] income > 38.9841: n = 22
|   |   |   |   |       (Intercept)
|   |   |   |   |         724.5455

Number of inner nodes:    26
Number of terminal nodes: 27
Number of parameters per node: 1
Objective function (negative log-likelihood): 7923.849
> plot(tree2)
> #Q4
> glmtree(sat~income+group,data=dtable)
Generalized linear model tree (family: gaussian)

Model formula:
sat ~ 1 | income + group

Fitted party:
[1] root
|   [2] group in 2
|   |   [3] income <= 78.92595
|   |   |   [4] income <= 73.90014
|   |   |   |   [5] income <= 72.76712: n = 31
|   |   |   |       (Intercept)
|   |   |   |         1355.806
|   |   |   |   [6] income > 72.76712
|   |   |   |   |   [7] income <= 73.47808: n = 13
|   |   |   |   |       (Intercept)
|   |   |   |   |         1288.462
|   |   |   |   |   [8] income > 73.47808: n = 10
|   |   |   |   |       (Intercept)
|   |   |   |   |             1338
|   |   |   [9] income > 73.90014
|   |   |   |   [10] income <= 75.53141: n = 42
|   |   |   |       (Intercept)
|   |   |   |         1278.571
|   |   |   |   [11] income > 75.53141: n = 121
|   |   |   |       (Intercept)
|   |   |   |         1244.215
|   |   |  [12] income > 78.92595
|   |   |   |   [13] income <= 83.33764
|   |   |   |   |   [14] income <= 81.03089: n = 93
|   |   |   |   |       (Intercept)
|   |   |   |   |          1200.43
|   |   |   |   |   [15] income > 81.03089: n = 77
|   |   |   |   |       (Intercept)
|   |   |   |   |         1161.429
|   |   |   [16] income > 83.33764
|   |   |   |   [17] income <= 87.82086: n = 88
|   |   |   |       (Intercept)
|   |   |   |         1114.318
```

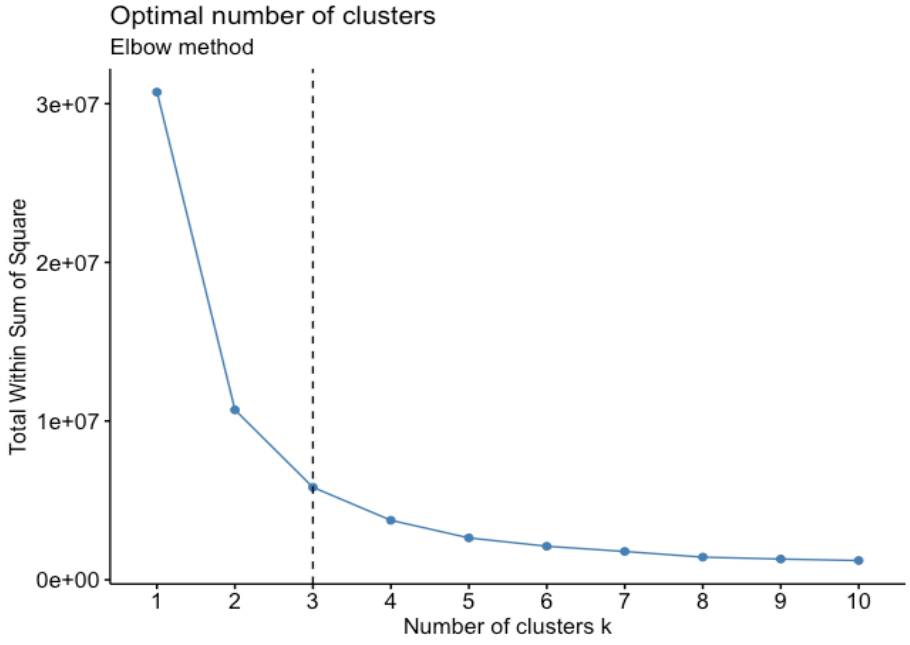
```
|   |   |   [18] income > 87.82086
|   |   |   |   [19] income <= 88.83985: n = 12
|   |   |   |       (Intercept)
|   |   |   |         1056.667
|   |   |   |   [20] income > 88.83985: n = 13
|   |   |   |       (Intercept)
|   |   |   |         983.8462
|   [21] group in 1, 3
|   |   [22] group in 1
|   |   |   [23] income <= 49.76334
|   |   |   |   [24] income <= 45.7473
|   |   |   |   |   [25] income <= 43.47638: n = 43
|   |   |   |   |       (Intercept)
|   |   |   |   |         1276.279
|   |   |   |   |   [26] income > 43.47638: n = 65
|   |   |   |   |       (Intercept)
|   |   |   |   |         1208.769
|   |   |   |   [27] income > 45.7473
|   |   |   |   |   [28] income <= 48.12029: n = 85
|   |   |   |   |       (Intercept)
|   |   |   |   |         1168.471
|   |   |   |   |   [29] income > 48.12029: n = 71
|   |   |   |   |       (Intercept)
|   |   |   |   |         1122.676
|   |   |   [30] income > 49.76334
|   |   |   |   [31] income <= 55.16208
|   |   |   |   |   [32] income <= 52.39517: n = 94
|   |   |   |   |       (Intercept)
|   |   |   |   |         1076.915
|   |   |   |   |   [33] income > 52.39517: n = 59
|   |   |   |   |       (Intercept)
|   |   |   |   |         1027.119
|   |   |   |   [34] income > 55.16208
|   |   |   |   |   [35] income <= 59.26473: n = 65
|   |   |   |   |       (Intercept)
|   |   |   |   |         972.9231
|   |   |   |   |   [36] income > 59.26473: n = 18
|   |   |   |   |       (Intercept)
|   |   |   |   |         847.7778
|   |   [37] group in 3
|   |   |   [38] income <= 31.89534
|   |   |   |   [39] income <= 26.82913
|   |   |   |   |   [40] income <= 19.76067: n = 12
|   |   |   |   |       (Intercept)
|   |   |   |   |         1308.333
|   |   |   |   |   [41] income > 19.76067
|   |   |   |   |   |   [42] income <= 25.03478: n = 53
|   |   |   |   |   |       (Intercept)
|   |   |   |   |   |         1158.302
|   |   |   |   |   |   [43] income > 25.03478: n = 52
|   |   |   |   |   |       (Intercept)
|   |   |   |   |   |         1098.846
|   |   |   |   [44] income > 26.82913
|   |   |   |   |   [45] income <= 29.59926: n = 105
|   |   |   |   |       (Intercept)
```

```
|   |   |   |   |         1043.048
|   |   |   |   |   [46] income > 29.59926: n = 92
|   |   |   |   |       (Intercept)
|   |   |   |   |         991.1957
|   |   |   [47] income > 31.89534
|   |   |   |   [48] income <= 36.12898
|   |   |   |   |   [49] income <= 33.79417: n = 68
|   |   |   |   |       (Intercept)
|   |   |   |   |         928.9706
|   |   |   |   |   [50] income > 33.79417: n = 60
|   |   |   |   |       (Intercept)
|   |   |   |   |         875.8333
|   |   |   |   [51] income > 36.12898
|   |   |   |   |   [52] income <= 38.9841: n = 36
|   |   |   |   |       (Intercept)
|   |   |   |   |         816.6667
|   |   |   |   |   [53] income > 38.9841: n = 22
|   |   |   |   |       (Intercept)
|   |   |   |   |         724.5455
```

Number of inner nodes: 26
Number of terminal nodes: 27
Number of parameters per node: 1
Objective function (negative log-likelihood): 7923.849

According to the data generation process, we can see that the tree has been divided first by group then by income.
The 1st group has been divided into 8 nodes, 2st group into 10 nodes 3rd group into 9 nodes.
Specifically, in group 1 the node having income between 52.39517 and 55.162 has the most members n= 94.
In group 2 the node having income > 75.53 has the most members n=121
In group 3 the node having income between 26.83 and 29.6 has the most members n=105

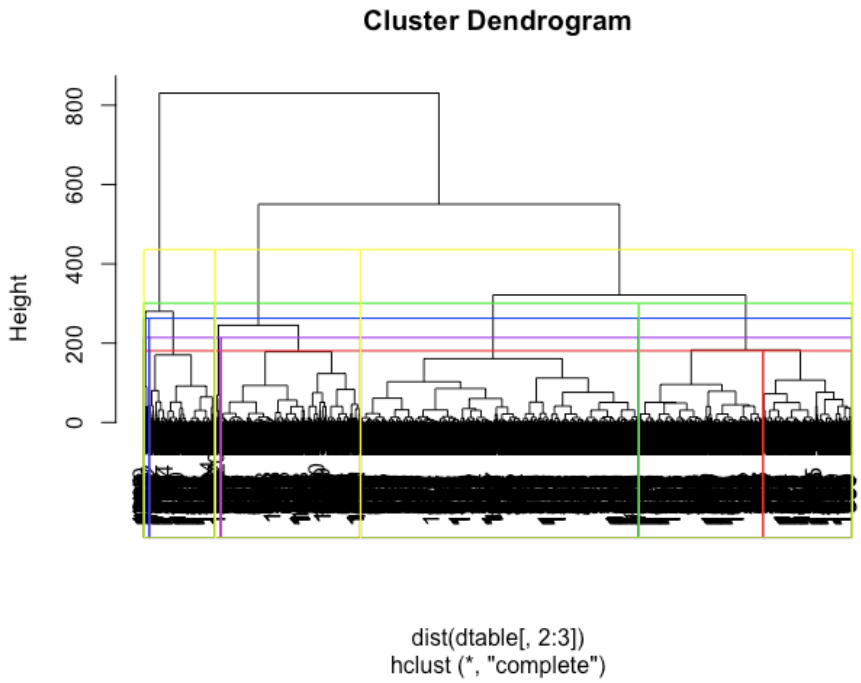
5. The the optimal number of groups here is 3.



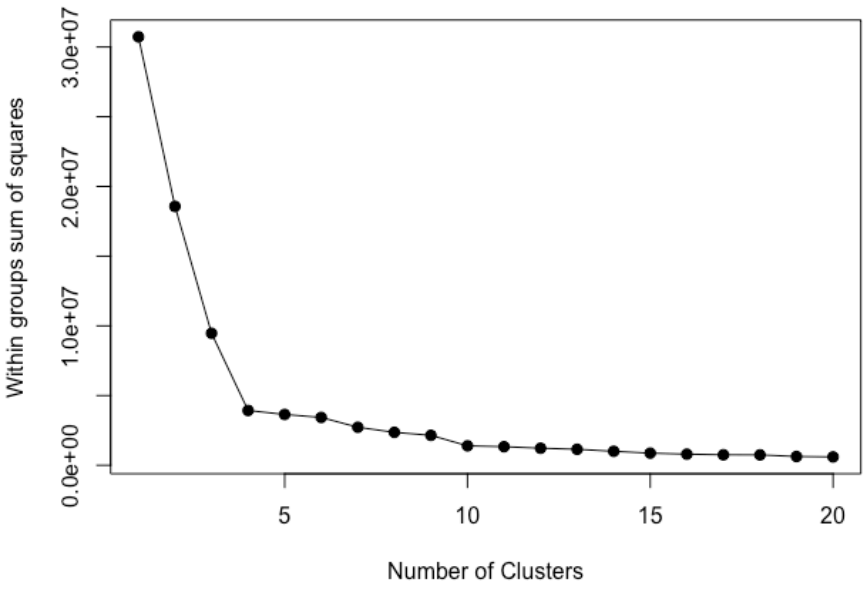
Correct means:

	sat	income
1	903.8527	41.84156
2	1243.0258	64.03255
3	1086.0165	50.51359

6.



We don't usually correctly identify the cluster of the data. In this case, k-mean give us the optimal number of cluster is 3 and hierarchical method gives us the optimal number of cluster is 4. In this case, k-mean is preferable. Although hierarchical is flexible but cannot be used on large dataset and it's more difficult to interpret the results compared to k-means method.



The optimal number of cluster here is 4.

7. Pooled OLS model:

```
Call:
lm(formula = sat ~ income, data = newdtable)

Residuals:
    Min      1Q  Median      3Q     Max
-452.84  -81.64    7.67   88.71  440.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 950.8914     9.1279  104.17   <2e-16 ***
income        2.7923     0.1593   17.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.1 on 1498 degrees of freedom
Multiple R-squared:  0.1703,  Adjusted R-squared:  0.1697
F-statistic: 307.4 on 1 and 1498 DF,  p-value: < 2.2e-16
```

Fixed-effects model using income and k-mean clustering:

```
Call:
lm(formula = sat ~ income + clusterNum - 1, data = newdtable)

Residuals:
     Min       1Q   Median       3Q      Max
-291.082  -39.449    0.443   42.542  202.860

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
income      1.875e-01  8.169e-02   2.296   0.0218 *
clusterNum1 1.083e+03  4.814e+00 225.077   <2e-16 ***
clusterNum2 9.029e+02  4.566e+00 197.747   <2e-16 ***
clusterNum3 1.234e+03  5.907e+00 208.907   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.53 on 1496 degrees of freedom
Multiple R-squared:  0.9971,  Adjusted R-squared:  0.9971
F-statistic: 1.297e+05 on 4 and 1496 DF,  p-value: < 2.2e-16
```

Fixed-effects model using income and hierarchical clustering:

```
Call:
lm(formula = sat ~ income + hier - 1, data = newdtable)

Residuals:
     Min       1Q   Median       3Q      Max
-211.733  -37.068   -1.724   39.119  168.246

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
income 1.396e-01  6.591e-02   2.118   0.0343 *
hier1  1.140e+03  4.408e+00 258.698   <2e-16 ***
hier2  9.967e+02  3.608e+00 276.233   <2e-16 ***
hier3  1.272e+03  4.967e+00 256.055   <2e-16 ***
hier4  8.256e+02  4.665e+00 176.980   <2e-16 ***
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.67 on 1495 degrees of freedom
Multiple R-squared: 0.9982, Adjusted R-squared: 0.9982
F-statistic: 1.621e+05 on 5 and 1495 DF, p-value: < 2.2e-16

We are able to find the relationships we know exist from the data generating process. The coefficients signs and values in these models are much different from the ones that we got at the second part. All these models are significant and the 2 with-in modes have really high R-square, while the OLS model has very low R-sq.

8.
Pooled OLS model:
Call:
lm(formula = sat ~ income, data = newdtable)

Residuals:
   Min   1Q Median   3Q   Max
-452.84 -81.64   7.67   88.71 440.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 950.8914   9.1279 104.17   <2e-16 ***
income   2.7923   0.1593   17.53   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.1 on 1498 degrees of freedom
Multiple R-squared: 0.1703, Adjusted R-squared: 0.1697
F-statistic: 307.4 on 1 and 1498 DF, p-value: < 2.2e-16

Fixed-effects model using income and k-mean clustering:

Call:
lm(formula = sat ~ income + clusterNumIncome - 1, data = newdtable)

Residuals:
   Min   1Q Median   3Q   Max
-689.10 -33.72   4.27   40.37 516.99

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
income   -16.6650   0.4612 -36.14   <2e-16 ***
clusterNumIncome1 1920.2418   23.2610   82.55   <2e-16 ***
clusterNumIncome2 2529.5404   36.9849   68.39   <2e-16 ***
clusterNumIncome3 1509.2876   14.4373 104.54   <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.94 on 1496 degrees of freedom
Multiple R-squared: 0.994,   Adjusted R-squared: 0.994
F-statistic: 6.207e+04 on 4 and 1496 DF, p-value: < 2.2e-16

We are able to find the relationships we know exist from the data generating process. Although these 2 models are significant, the OLS model has very low R-sq while the fixed-effects model has very high R-sq.

9.

Pooled OLS model:

```
Call:
lm(formula = sat ~ income, data = newdtable)

Residuals:
   Min     1Q Median     3Q    Max
-452.84 -81.64   7.67  88.71 440.50

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 950.8914    9.1279      104.17   <2e-16 ***
income       2.7923     0.1593      17.53    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.1 on 1498 degrees of freedom
Multiple R-squared:  0.1703,  Adjusted R-squared:  0.1697
F-statistic: 307.4 on 1 and 1498 DF,  p-value: < 2.2e-16
```

Fixed-effects model using income and k-mean clustering:

```
Call:
lm(formula = sat ~ income + clusterNumScale - 1, data = newdtable)

Residuals:
    Min      1Q  Median      3Q     Max
-302.745  -57.212    4.527   58.124  253.949

Coefficients:
                       Estimate   Std. Error  t value   Pr(>|t|)
income                 -2.0567     0.2263     -9.088    <2e-16 ***
clusterNumScale1 1012.5670     9.6937    104.456   <2e-16 ***
clusterNumScale2 1366.1819    18.4311   74.124    <2e-16 ***
clusterNumScale3 1232.1075     9.8771    124.744   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.92 on 1496 degrees of freedom
Multiple R-squared:  0.9946,  Adjusted R-squared:  0.9945
F-statistic: 6.834e+04 on 4 and 1496 DF,  p-value: < 2.2e-16
```

These 2 models are significant. The Pooled OLS model is still the same with or without scaling, while the fixed-effect model using k-means generate different coefficients with the same R-sq.