BUAN 6356.501 - Business Analytics with R (Spring 2019)
Problem Set 5

Question 1

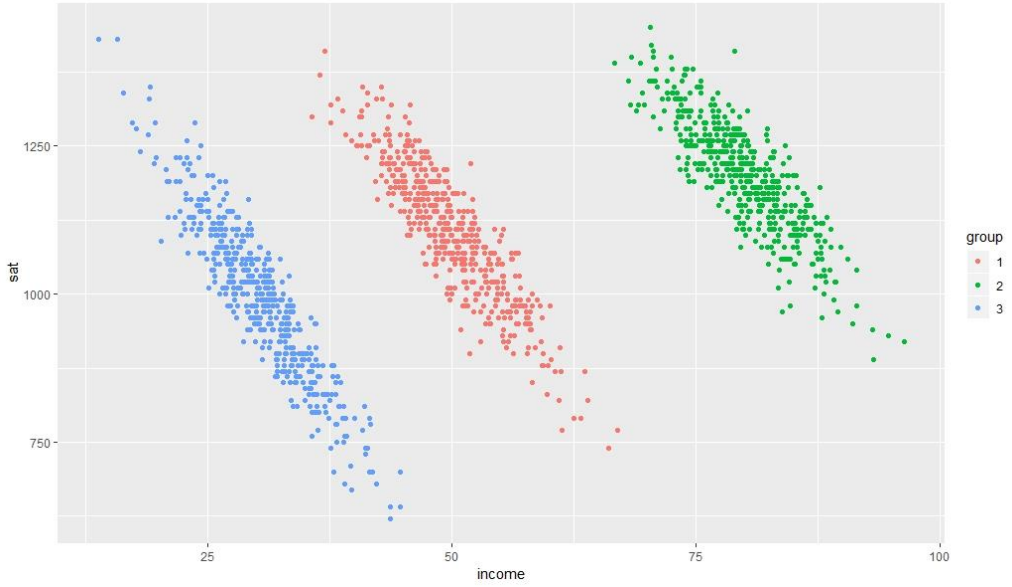
The variables *sat* and *income* are generated from the following linear equations for 3 groups.

Group 1: $sat = -100z + 1100 + 50w$ and $income = 5z + 50$

Group 2: $sat = -80z + 1200 + 50w$ and $income = 5z + 80$

Group 3: $sat = -120z + 1000 + 50w$ and $income = 5z + 30$

where *z* and *w* are random variables from a normal distribution



Question 2

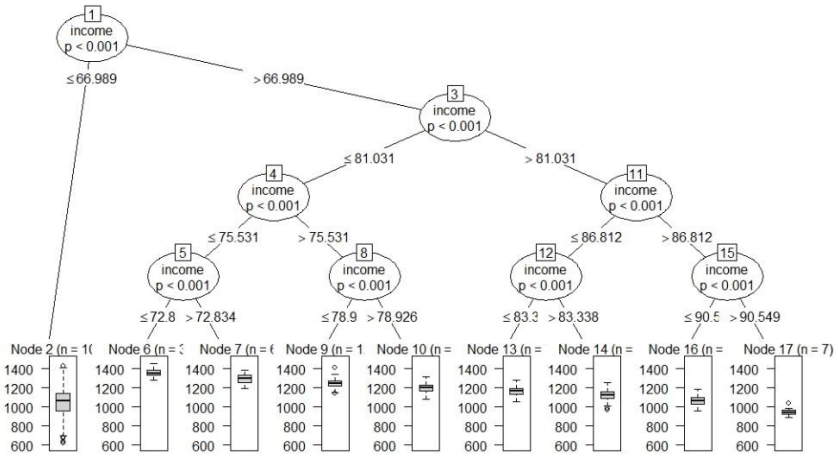
Pooled model does not take groups into account and generates a regression line considering all points as individual observations. Hence, the predicted line has a positive slope for income. But the fixed-effects model considers groups and predicts 3 lines having negative slopes for income. This can be confirmed by the similar results we get from running the individual models separately.

```
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
```

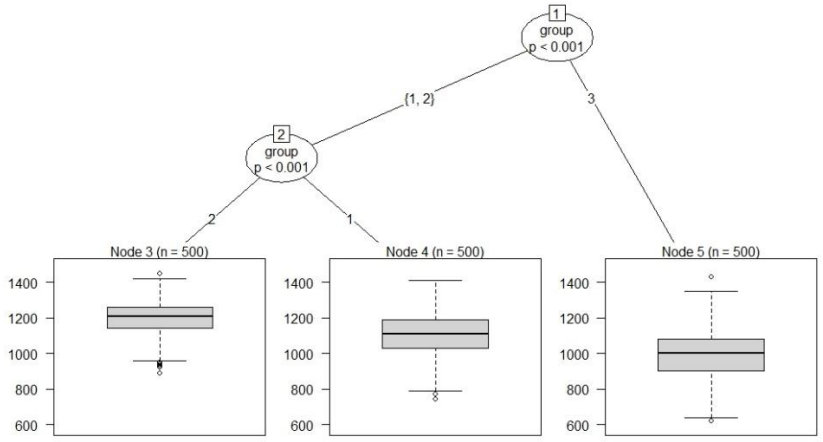
Question 3

We get below tree diagrams for 3 models, one with income, one with group and one with both.

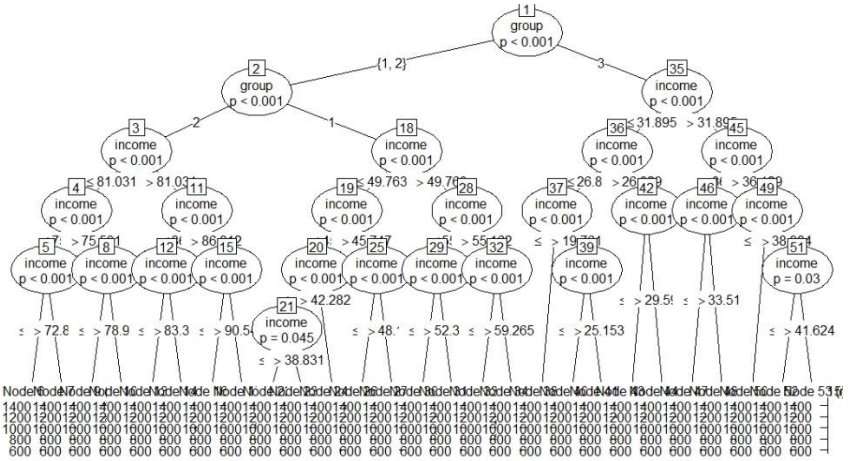
sat using only *income*



sat using only *group*

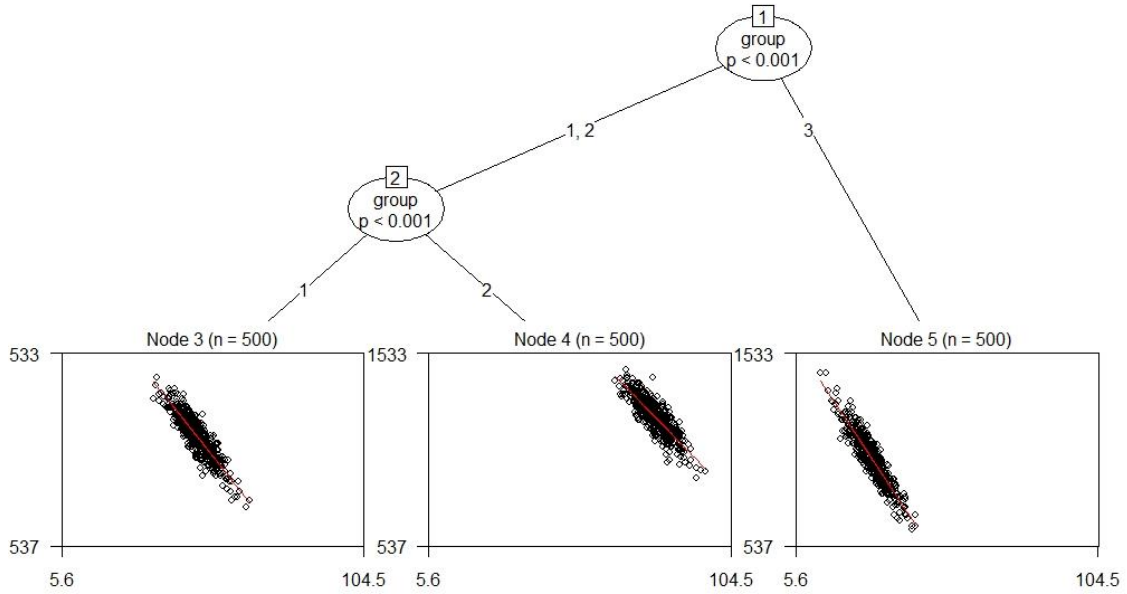


sat using *income* and *group*



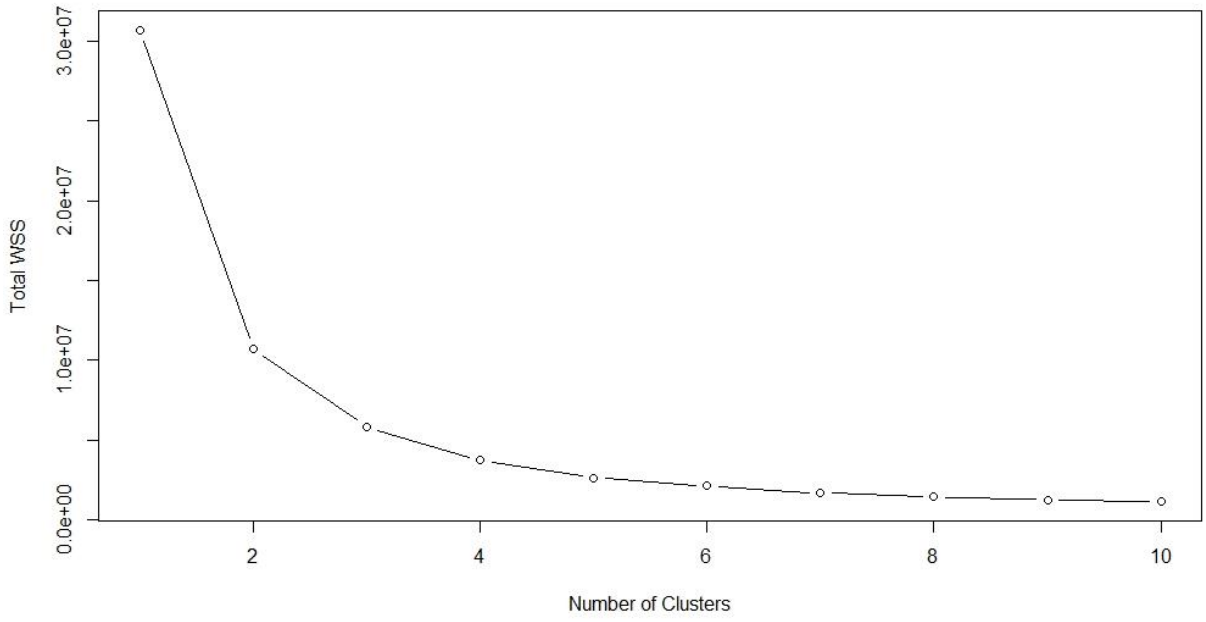
Question 4

We get the following tree diagram by running a glmtree model which is accurate as per the data.



Question 5

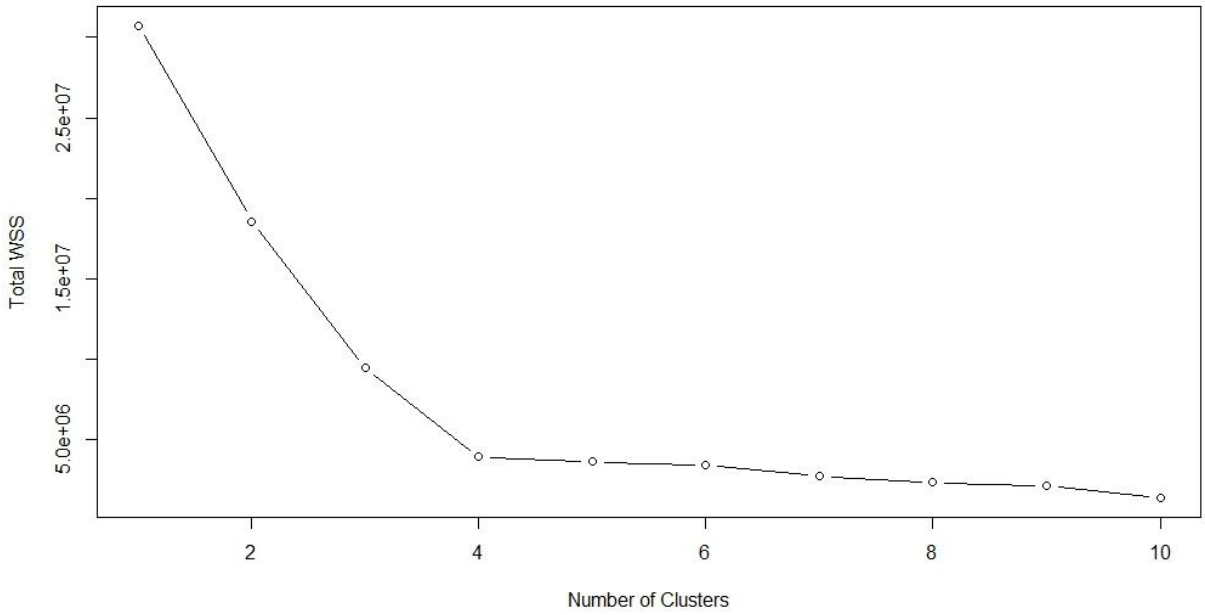
We get the optimal number of K-means clusters as 2 from the eigen values ratio and elbow plot.

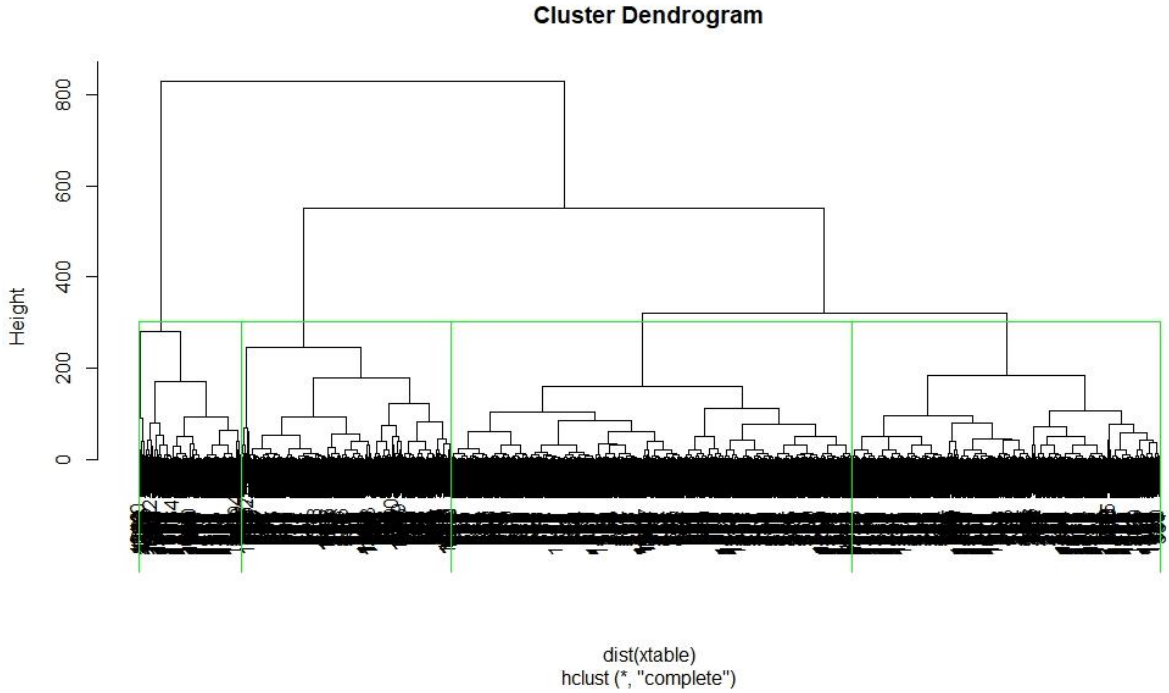


59.4% of group1 observations fall in cluster 1 and 40.6% of group1 observations fall in cluster 2.
89.2% of group2 observations fall in cluster 1 and 10.8% of group2 observations fall in cluster 2.
23.2% of group3 observations fall in cluster 1 and 76.8% of group3 observations fall in cluster 2.

Question 6

When we know that 3 groups exist, K-means gave 2 and failed to correctly identify the groups. But we get optimal number of hierarchical clusters as 4 from eigen values ratio and elbow plot. Both the models could not predict the correct number of groups as optimal number of clusters.





Question 7

The pooled model remains same as it would not consider any groups or clusters in the model. Due to incorrect number of clusters, even fixed-effects model shows positive slope for income.

K-means fixed-effects model

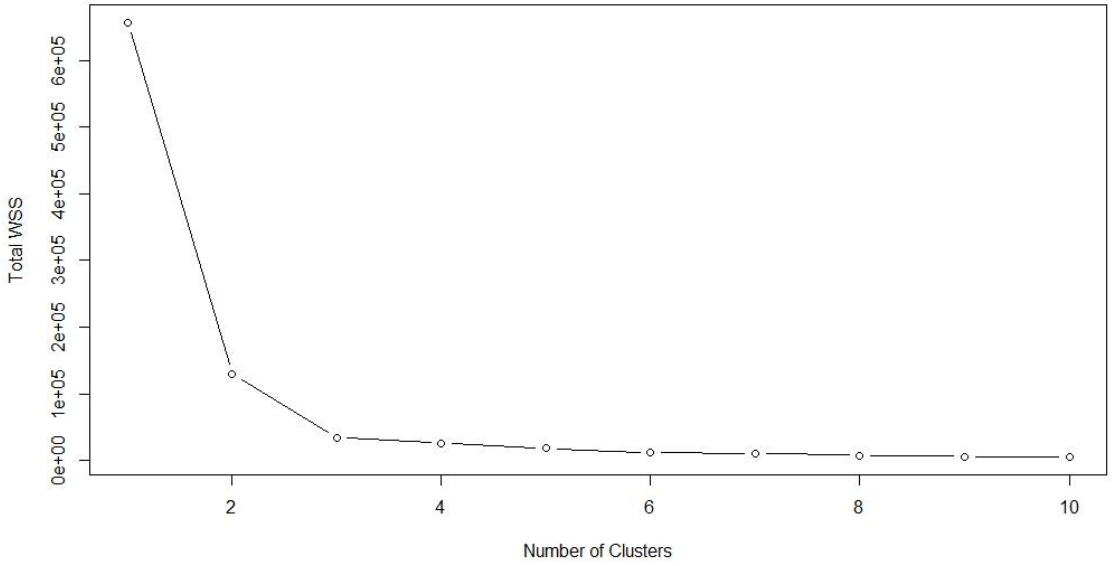
```
Call:
lm(formula = sat ~ income + kgroup - 1, data = xtable)

Residuals:
    Min      1Q  Median      3Q     Max
-346.37  -57.78    2.35   58.88  264.88

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
income     0.7354     0.1089   6.752 2.08e-11 ***
kgroup1 1154.9498     7.1458 161.626  < 2e-16 ***
kgroup2  934.2559     5.7478 162.541  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 81.12 on 1497 degrees of freedom
Multiple R-squared:  0.9947,    Adjusted R-squared:  0.9946
F-statistic: 9.294e+04 on 3 and 1497 DF,  p-value: < 2.2e-16
```

Hierarchy cluster fixed-effects model

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

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 *
hgroup1 1.140e+03  4.408e+00 258.698   <2e-16 ***
hgroup2 9.967e+02  3.608e+00 276.233   <2e-16 ***
hgroup3 1.272e+03  4.967e+00 256.055   <2e-16 ***
hgroup4 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
```

Question 8

When only single variable such as income is considered, we get optimal K-means clusters as 3. Cluster 1 has a 52% accuracy, Cluster 2 has a 71% accuracy and Cluster 3 has a 47% accuracy.



Question 9

When scaled data is considered for multiple variables, we get optimal K-means clusters as 3. Cluster 1 has a 74% accuracy, Cluster 2 has a 100% accuracy and Cluster 3 has a 66% accuracy. The accuracy of K-means model increased when scaled data is used. This could be the reason for the incorrect optimal number of clusters when we used multiple variables without scaling.