HW1

Xingwei Ji

6/15/2018

1. The total observation:

```
data <- readRDS("college_scorecard_2013.rds")
dim(data)
```

```
## [1] 3312   51
```

there are total of 3312 observations.

Total colleges:

```
length(data[,"name"])
```

```
## [1] 3312
```

there are 3312 different colleges.

2. number of features:

```
ncol(data)
```

```
## [1] 51
```

There are 51 different features.

How many of these are categorical:

```
factor_vector = names(data)[sapply(data, class) == "factor"]
logical_vector = names(data)[sapply(data, class) == "logical"]

length(factor_vector) + length(logical_vector)
```

```
## [1] 7
```

There are 7 catagerocal features.

How many are discrete:

```
table(sapply(data,class))
```

```
## 
## character    factor   integer   logical   numeric 
##         4         4        15         3        25
```

There are 15 descrete variables. There are also continuous and character features

3. Missing values:

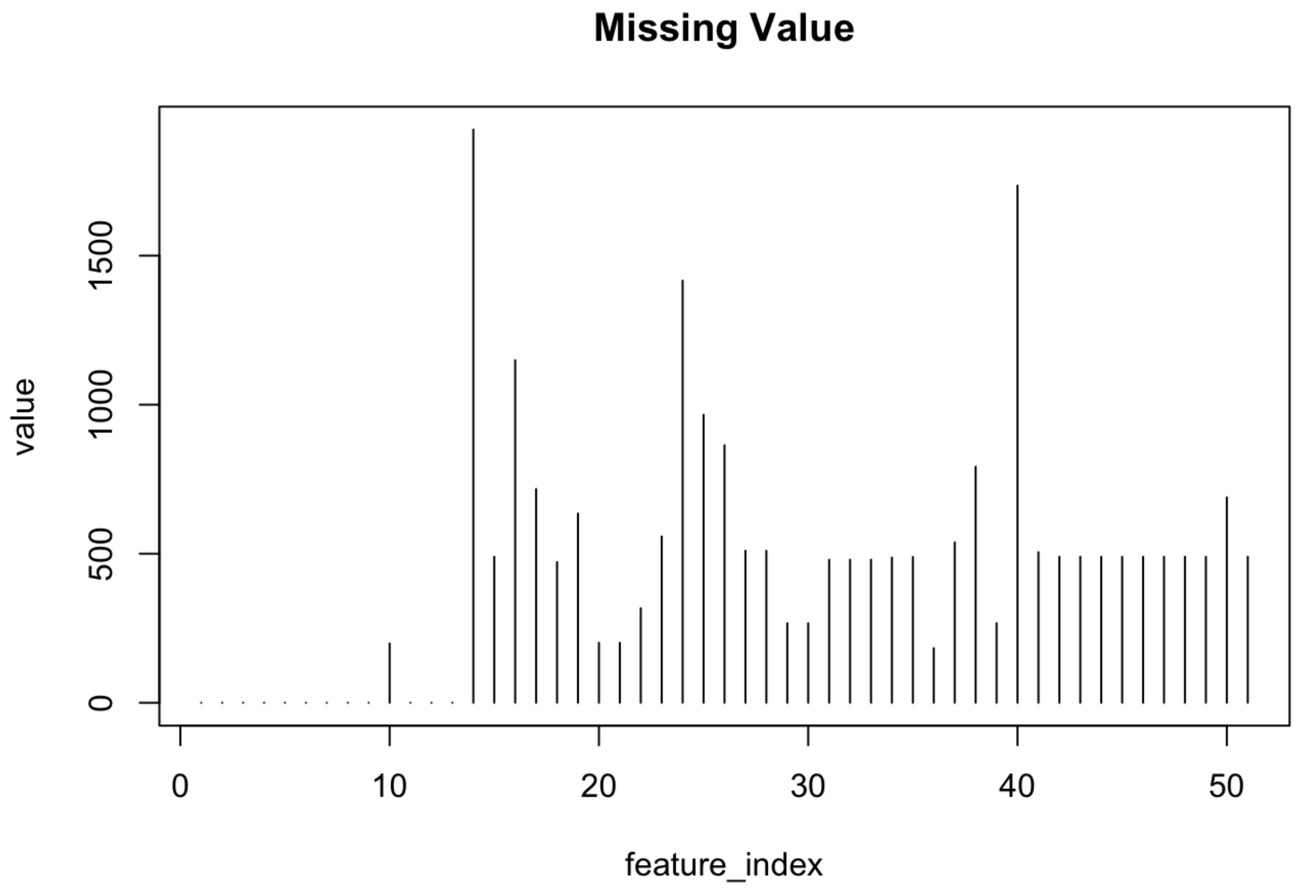
```
col_NAs = colSums(is.na(data))
sum(col_NAs)
```

```
## [1] 23197
```

```
names(data)[which.max(col_NAs)]
```

```
## [1] "avg_sat"
```

```
plot(col_NAs,type = "h",main = "Missing Value",xlab = "feature_index",ylab = "value
")
```



```
tail(sort(col_NAs))
```

```
## completion  retention    grad_pop  admission    veteran    avg_sat
##        864        966        1149       1416       1735       1923
```

We have total of 23197 missing values. The feature that has most missing values is “avg_sat”. We can see that those private information such as avg_sat and admission is mostly missing.

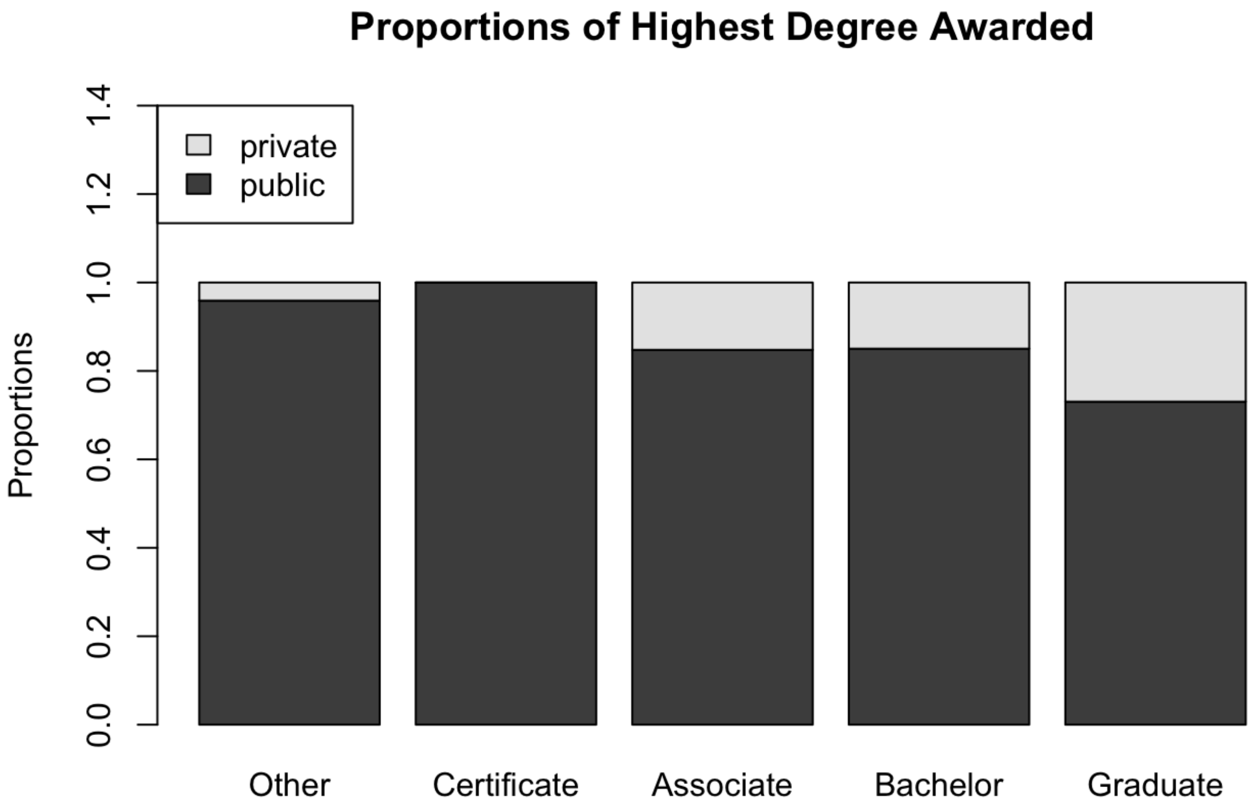
4. Public vs private colleges:

```
isPublic = factor(data$ownership == "Public",labels = c("Public", "Private"))
summary(isPublic)
```

```
##   Public Private
##     2596     716
```

There are 716 public schools, 2596 pravite schools.

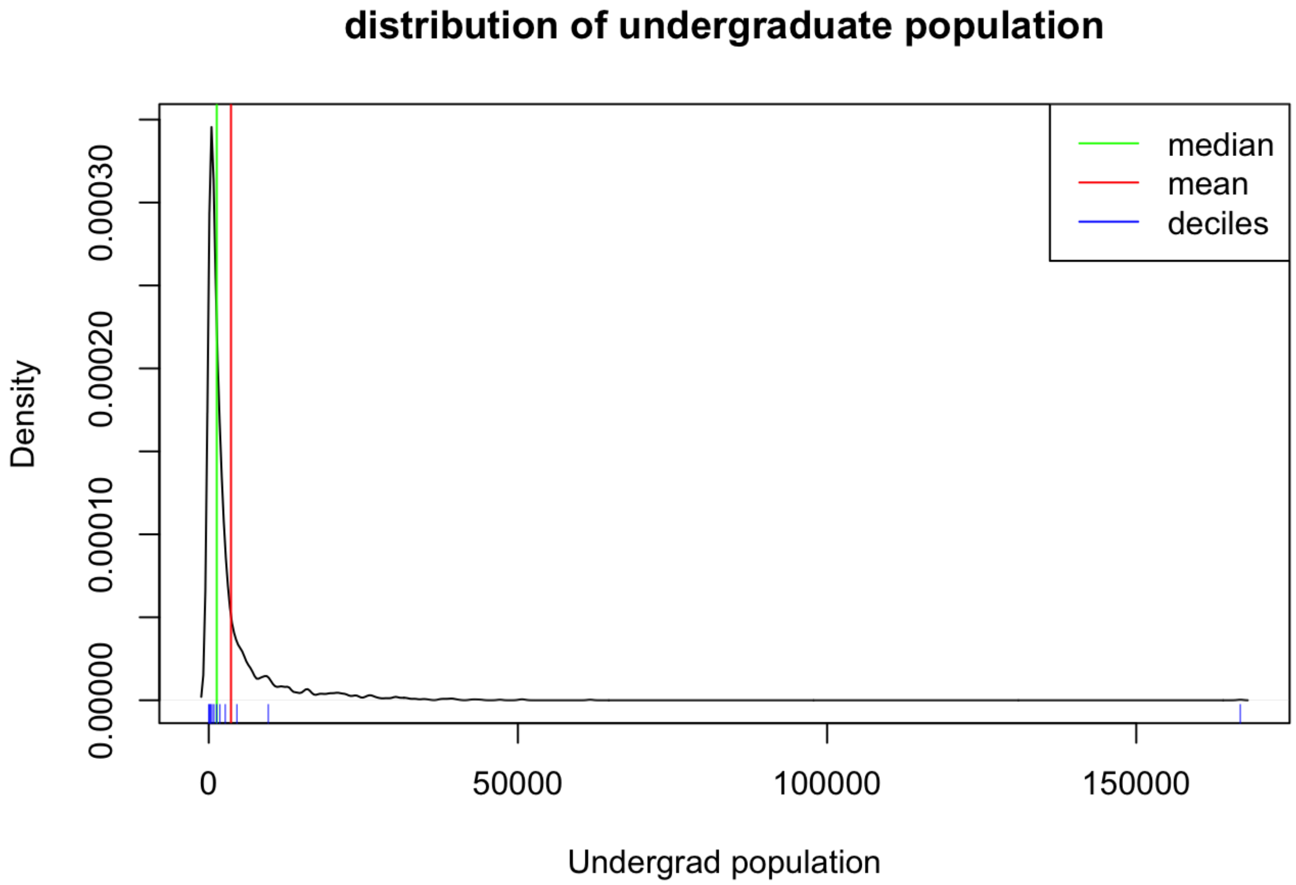
```
p_v_d_table = table(isPublic,data$highest_degree)
barplot(prop.table(p_v_d_table,margin=2),ylim=c(0,1.4),
        main="Proportions of Highest Degree Awarded",y="Proportions",
        legend=c("public","private"), args.legend = list(x="topleft"))
```



Public schools always have more propotions of highest degree awarded.

5. undergraduate population:

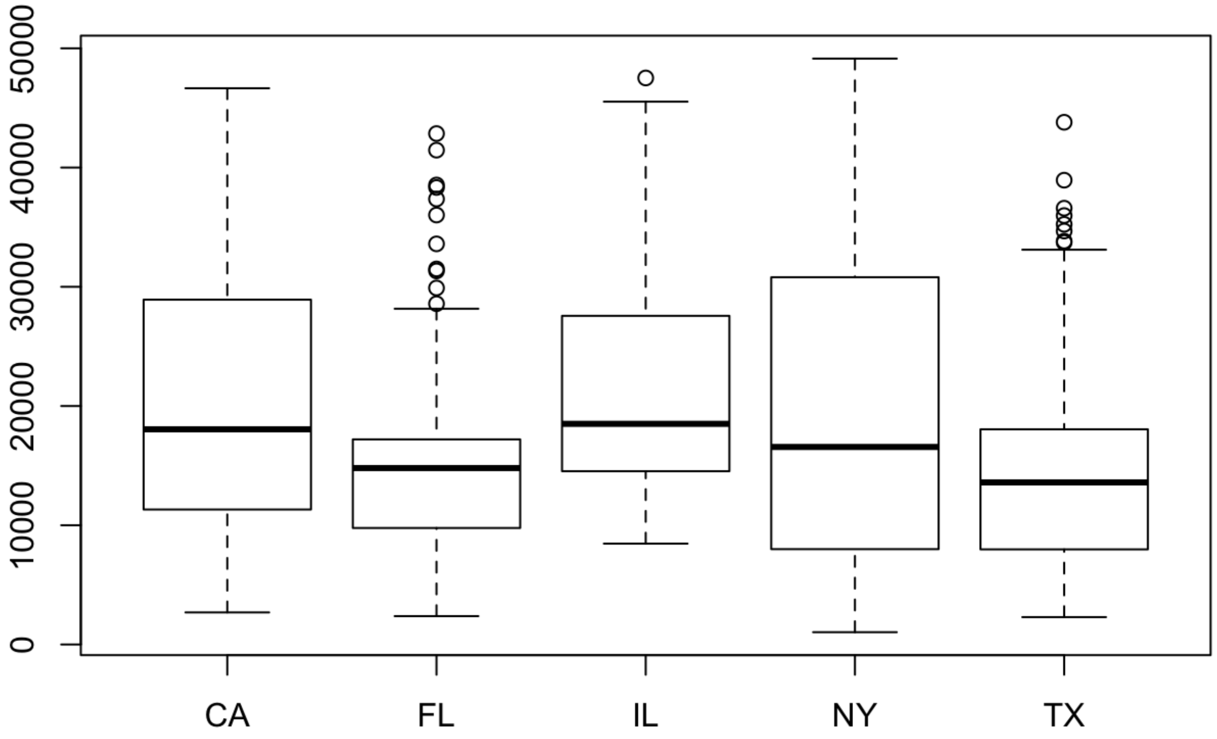
```
Un_pop_mean = mean(data$undergrad_pop,na.rm = TRUE)
Un_pop_median = median(data$undergrad_pop, na.rm = TRUE)
Un_pop_decile = quantile(data$undergrad_pop,seq(0,1,by = 0.1),na.rm = TRUE)
plot(density(data$undergrad_pop,na.rm = TRUE),main = "distribution of undergraduate
population", xlab = "Undergrad population")
abline(v = Un_pop_median, col= "green")
abline(v = Un_pop_mean,col = "red")
rug(x = Un_pop_decile,col = "blue")
#adding legend:
legend("topright",legend = c("median", "mean", "deciles"),lty = c("solid","solid","
solid"),col = c("green","red","blue"))
```



we can see that there is an oulier at population > 15000

6. Tuition:

```
#The five most populous states are:
#California, Texas, New York, Illinois, Florida.
pop_tuition = subset(data,(data$state == "CA" | data$state == "TX"| data$state == "
NY" |data$state == "IL" |data$state == "FL"),select = c("tuition","state"))
pop_tuition = droplevels(pop_tuition)
boxplot(tuition~state,data = pop_tuition)
```



we can see that CA, IL and NY have relatively high tuitions. However, there are some colleges in FL and TX having over 30000 tuitions fees as well.

7.

a.

```
rbysat = data[order(-data$avg_sat),]
rbysat[1,"name"]
```

```
## [1] "California Institute of Technology"
```

It’s California Institute of Technology
b.

```
sbunp = data[order(-data$undergrad_pop),]
sbunp[1,"open_admissions"]
```

```
## [1] TRUE
```

Yes, it has open admissions
c.

```
tf = data$ownership == "Public"
pubschool = data[tf,]
sbafi_pub = pubschool[order(pubschool$avg_family_inc),]
sbafi_pub[1,"zip"]
```

```
## [1] "11101"
```

The zip code us 11101.
d.

```
sbunp[1,"name"]
```

```
## [1] "University of Phoenix-Online Campus"
```

```
data[which.max(data$grad_pop),"name"]
```

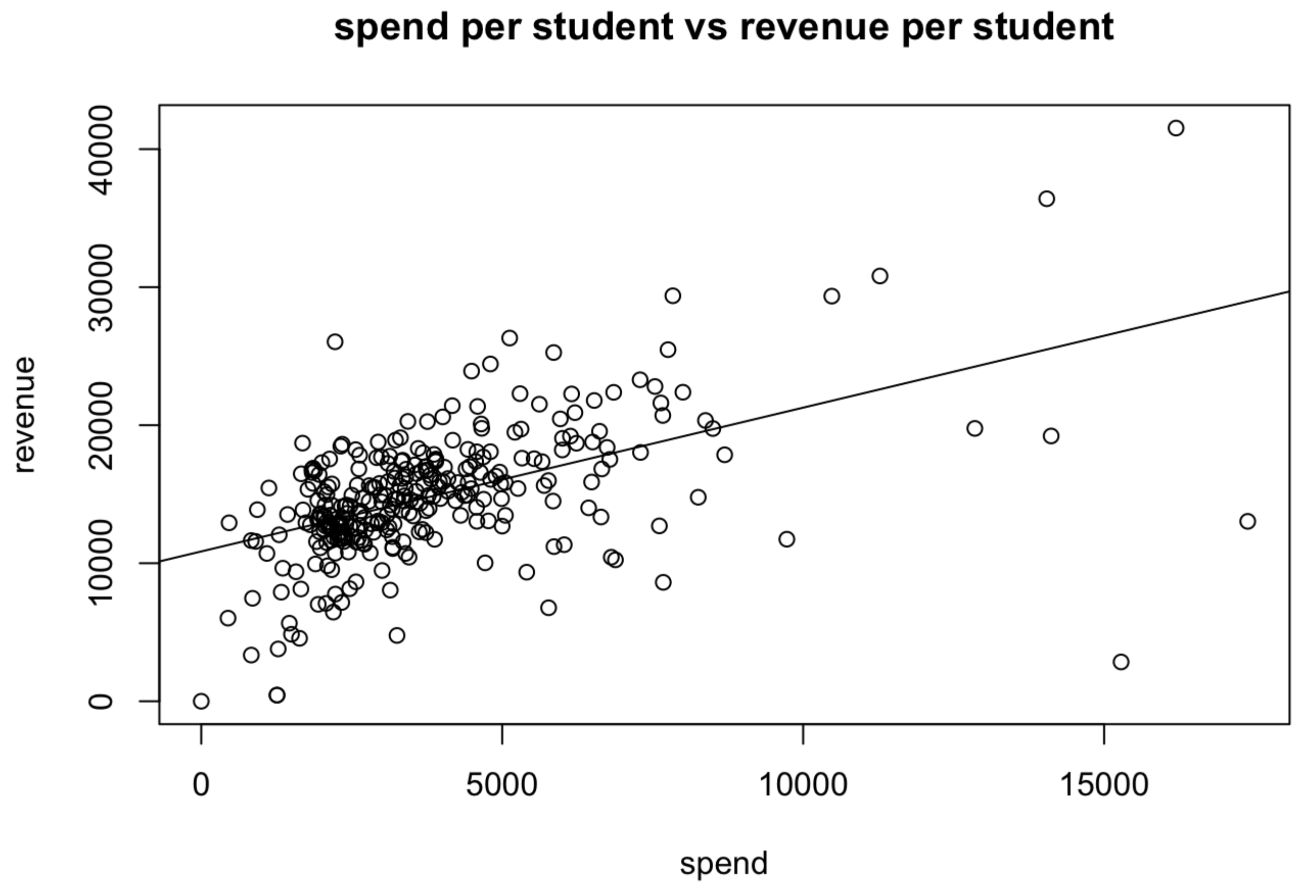
```
## [1] "Walden University"
```

The name from b is University of Phoenix-Online Campus, the school with largest graduate population is Walden University. It’s two differen school. So, no, it doesn’t have largest graduate population.

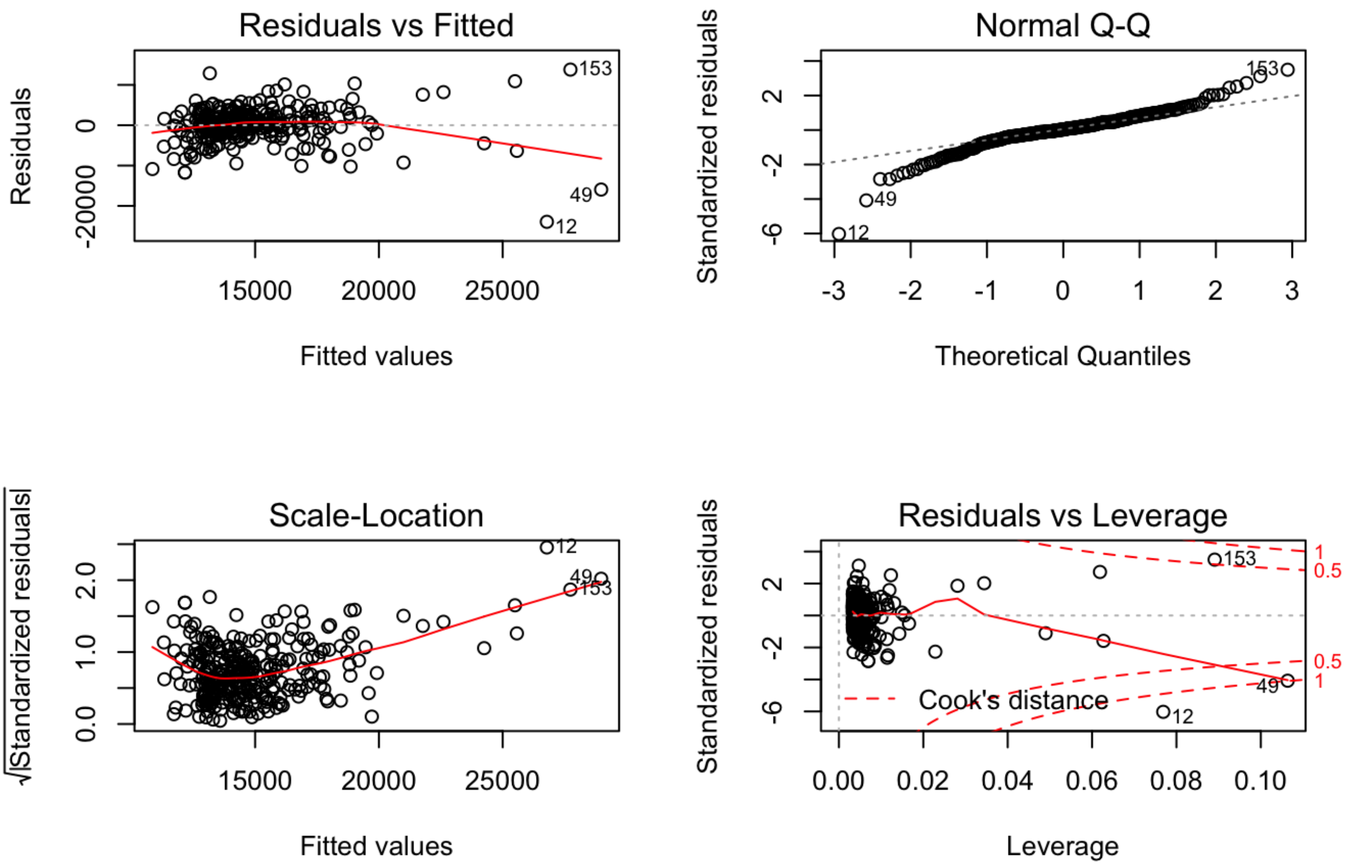
8. schools that are for-profit in ownership and issue Bachelor’s degrees as their primary_degree

a.

```
data8 = data[which(data$ownership == "For Profit" & data$primary_degree == "Bachelor"),]
model = lm(data8$revenue_per_student~data8$spend_per_student)
plot(data8$spend_per_student,data8$revenue_per_student, xlab = "spend",ylab = "revenue",main = "spend per student vs revenue per student")
abline(model)
```



```
par(mfrow = c(2,2))
plot(model)
```



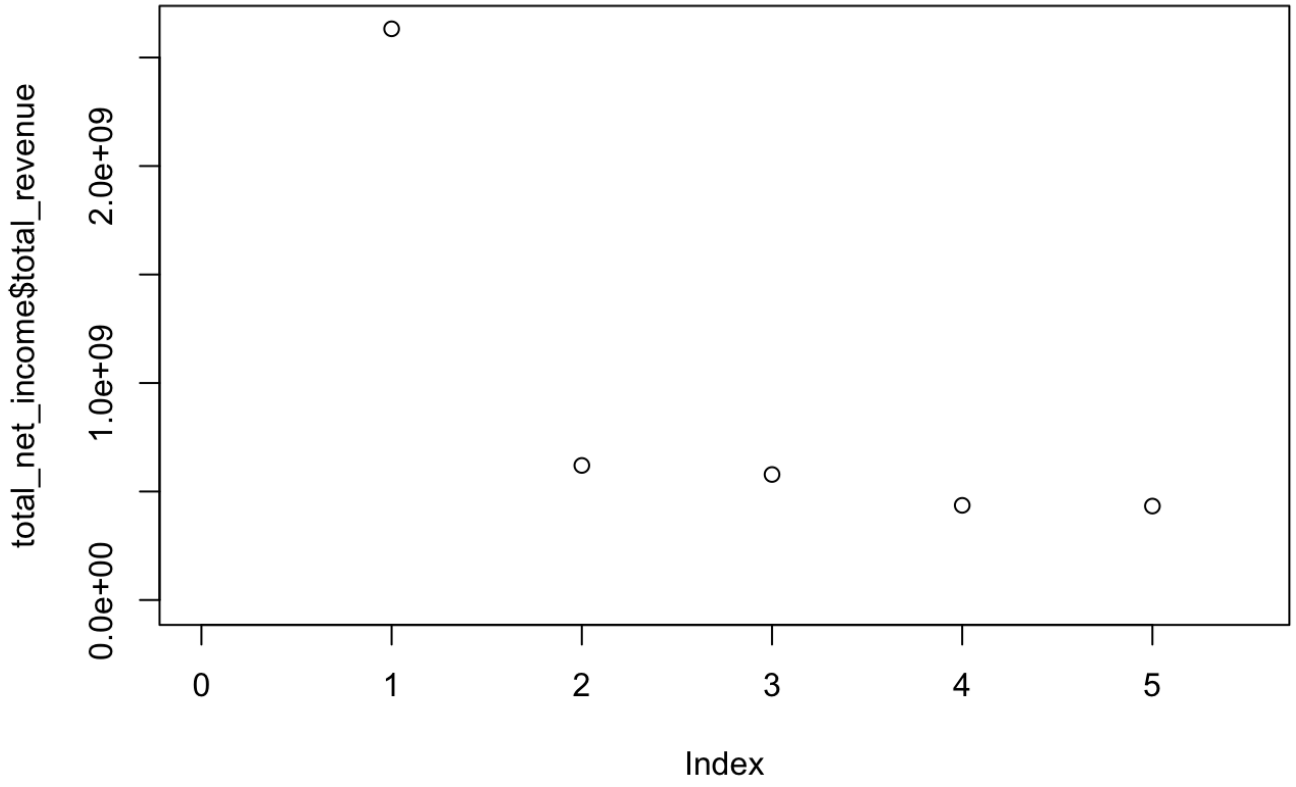
We can see that spend and revenue have roughly linear relationship. When we usually fit the data with linear model, we assume that normality and equal variance hold. Apparently this data violates equal variance and normality.

b. total_net_income net income should be (revenue - spend) * total quantity

```
revenue_per_student = data8$revenue_per_student - data8$spend_per_student
data8[is.na(data8$grad_pop),"grad_pop"] = 0
data8[is.na(data8$undergrad_pop),"undergrad_pop"] = 0
total_student = data8$undergrad_pop + data8$grad_pop
data8$total_revenue = revenue_per_student * total_student
total_net_income = data8[order(-data8$total_revenue),]
total_net_income[1:5,"name"]
```

```
## [1] "University of Phoenix-Online Campus"
## [2] "Ashford University"
## [3] "Capella University"
## [4] "Grand Canyon University"
## [5] "Kaplan University-Davenport Campus"
```

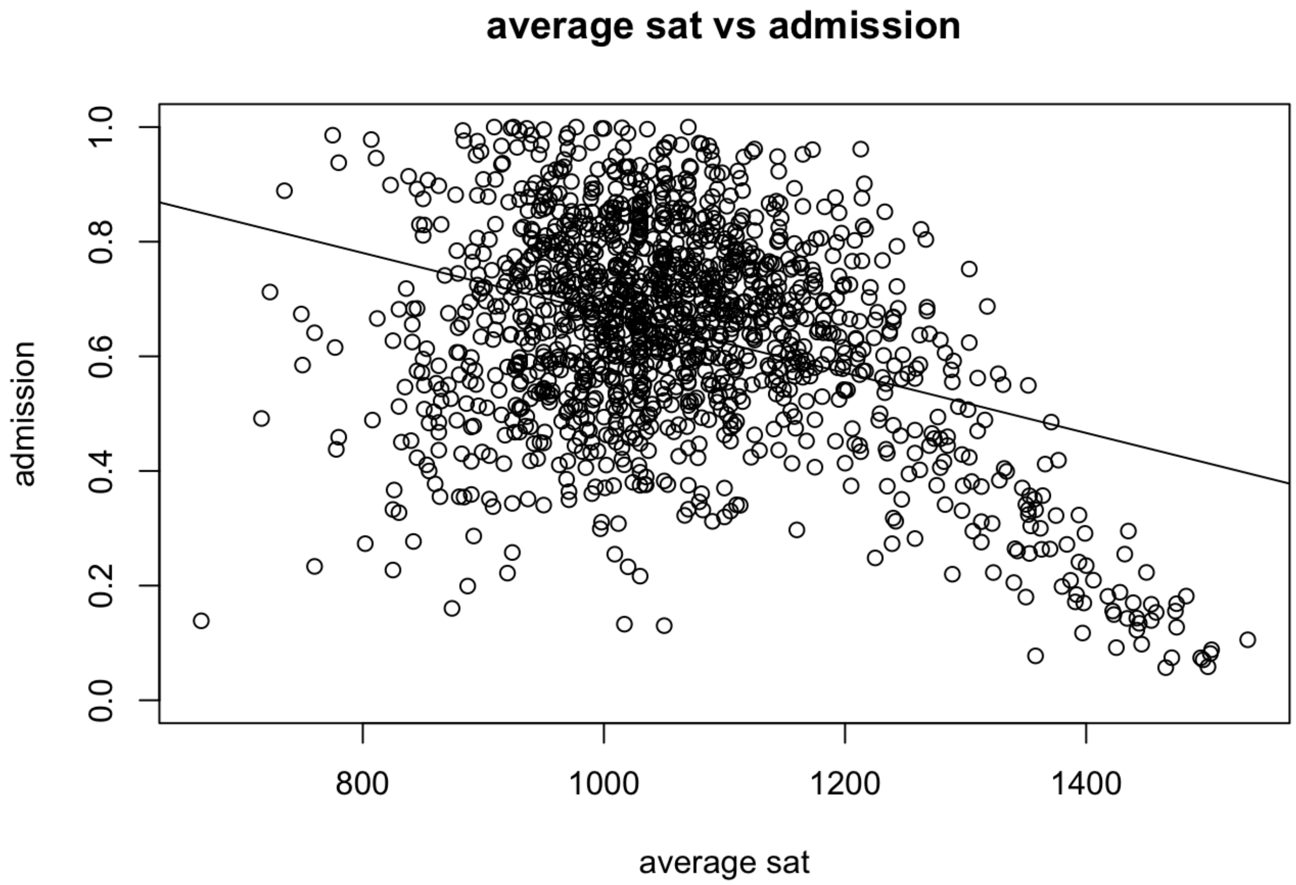
```
plot(total_net_income$total_revenue, xlim = c(0,5.5))
```



9. relationship between avg_sat and admission for all schools

a.

```
model2 = lm(data$admission~data$avg_sat)
plot(data$avg_sat,data$admission,xlab = "average sat", ylab = "admission",main = "a
verage sat vs admission")
abline(model2)
```

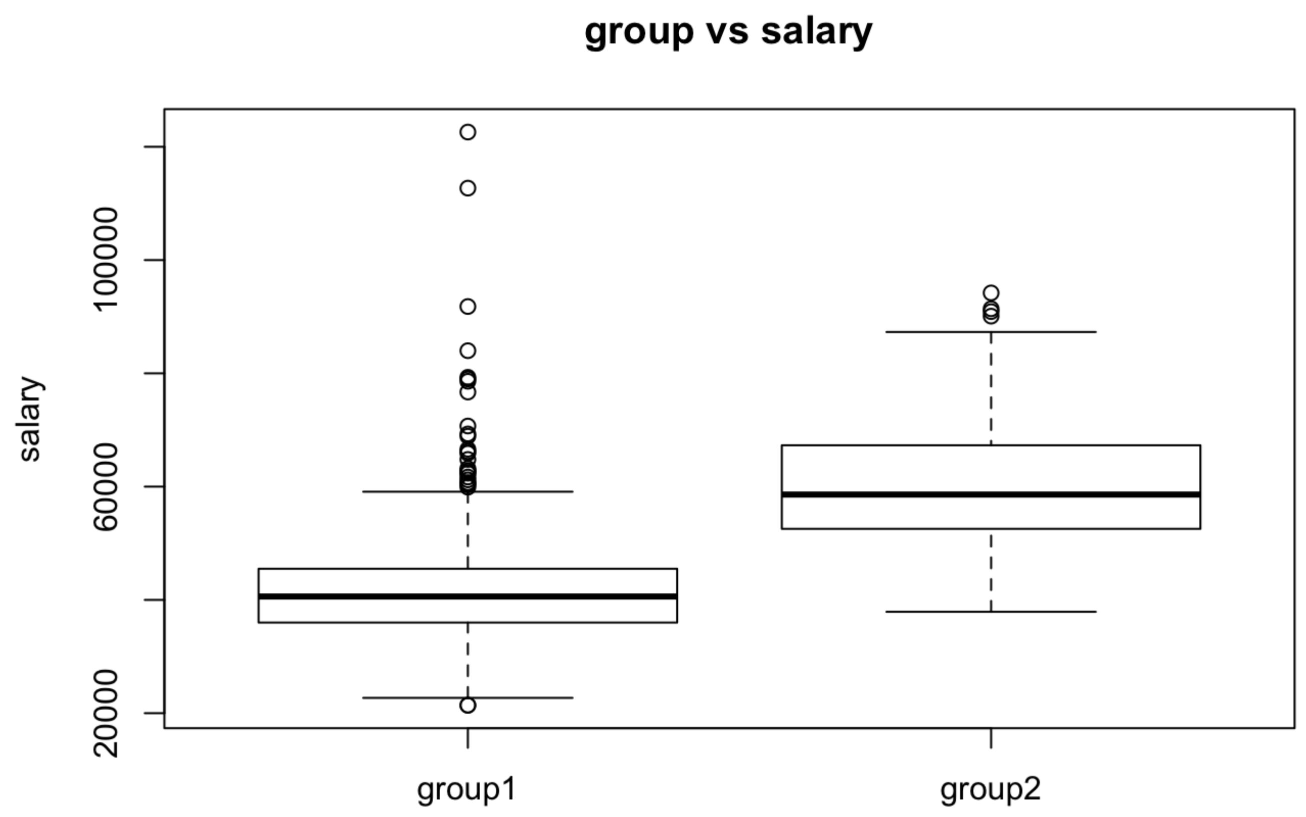


We can divide group1 with sat < 1200 and admisison > 0.6 group2 as sat > 1200 and admission < 0.6

```
data9 = data
group1_ind = which(data9$avg_sat <= 1200)
group2_ind = which(data9$avg_sat > 1200 & data9$admission < 0.6)
data9[group1_ind,"group"] = "group1"
data9[group2_ind,"group"] = "group2"
```

b. (a). med_10yr_salary

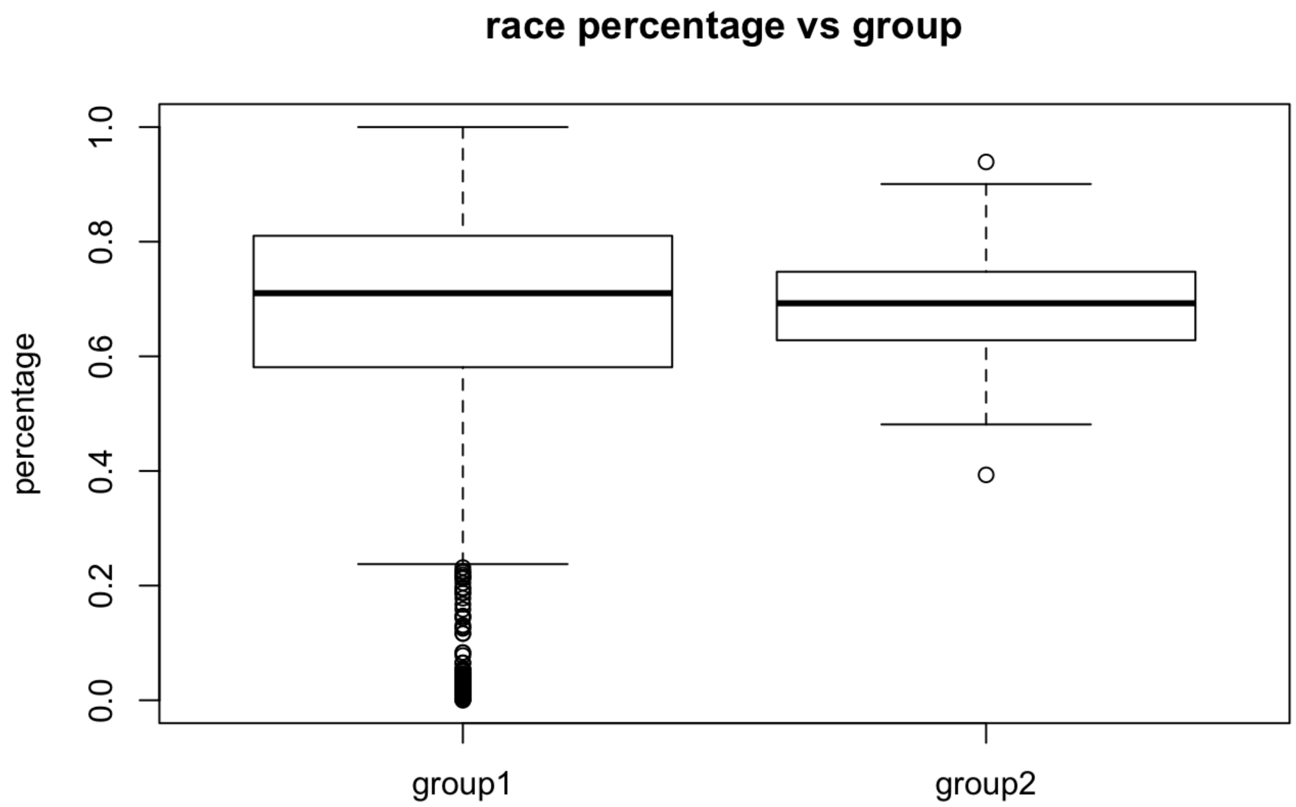
```
boxplot(data9$med_10yr_salary~data9$group, main = "group vs salary",ylab = "salary"
)
```



Group2 with higher sat has generally higher salary than group 1

(b). The percentage of race_white and race_asian combined

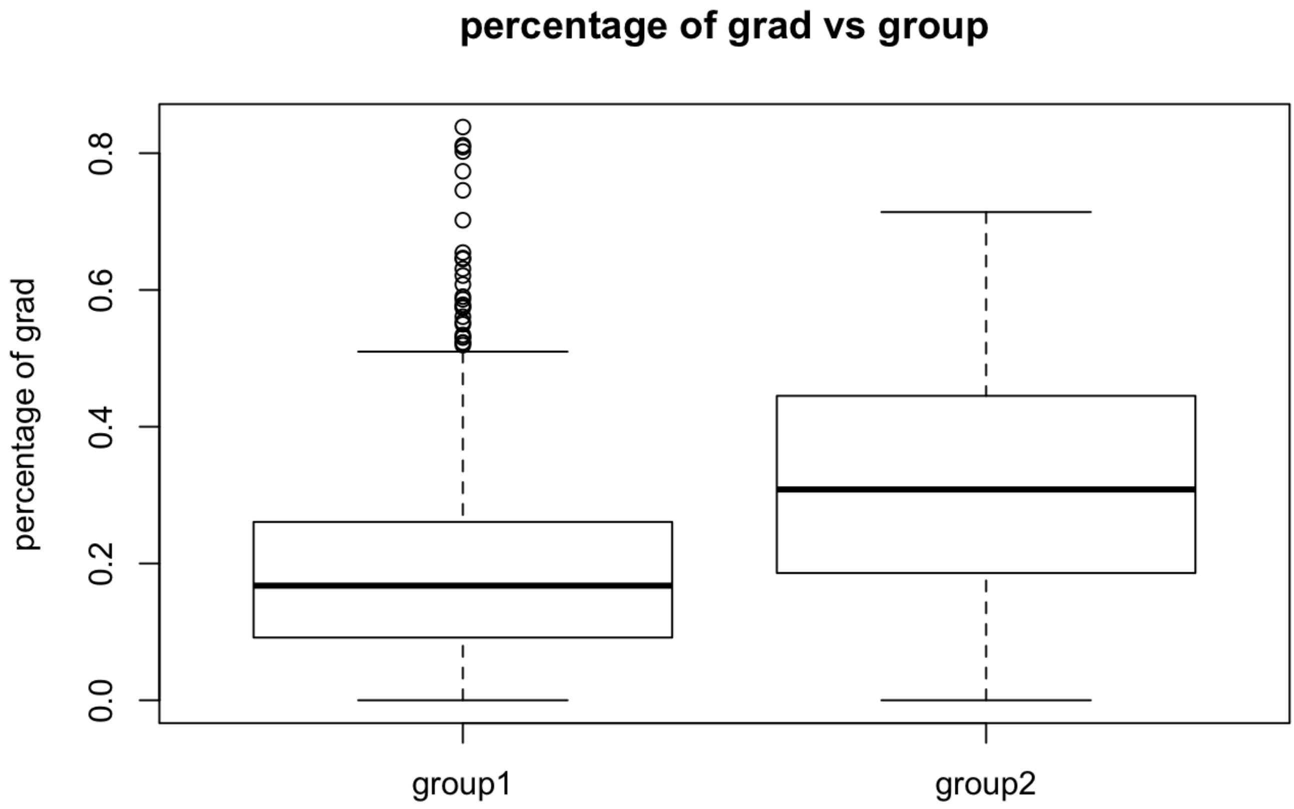
```
data9$all_race = data9$race_asian + data9$race_white
boxplot(data9$all_race~data9$group,main = "race percentage vs group", ylab = "perce
ntage")
```



Group1 with lower sat and admission has higher percentage of race_white and race_asian combined. It is also more spread out compared with group2 which has higher sat and lower admission rate.

(c). The percentage of graduate students enrolled at a university

```
data9$percetage_of_grad = data9$grad_pop/(data9$undergrad_pop + data9$grad_pop)
boxplot(data9$percetage_of_grad~data9$group,main = "percentage of grad vs group", y
lab = "percentage of grad")
```



group2 with higher sat and lower admission rate has higher percentage of graduate students in average and more spread than the group1.

c. open_admission, main campus, ownership, more than 1 branch or not

```
table(data9$open_admissions,data9$group)
```

```
##
##          group1 group2
##   FALSE    1209    136
##   TRUE        0      0
```

```
table(data9$main_campus,data9$group)
```

```
##
##          group1 group2
##   FALSE      45      0
##   TRUE     1164    136
```

```
table(data9$ownership,data9$group)
```

```
##
##               group1 group2
##   Public         477     30
##   Nonprofit      723    106
##   For Profit       9      0
```

```
more_than_1_branch_index = which(data9$branches > 1)
less_than_1_branch_index = which(data9$branches <= 1)
data9[more_than_1_branch_index,"mt1b"] = "More than 1"
data9[less_than_1_branch_index,"mt1b"] = "Less than 1"
table(data9$mt1b,data9$group)
```

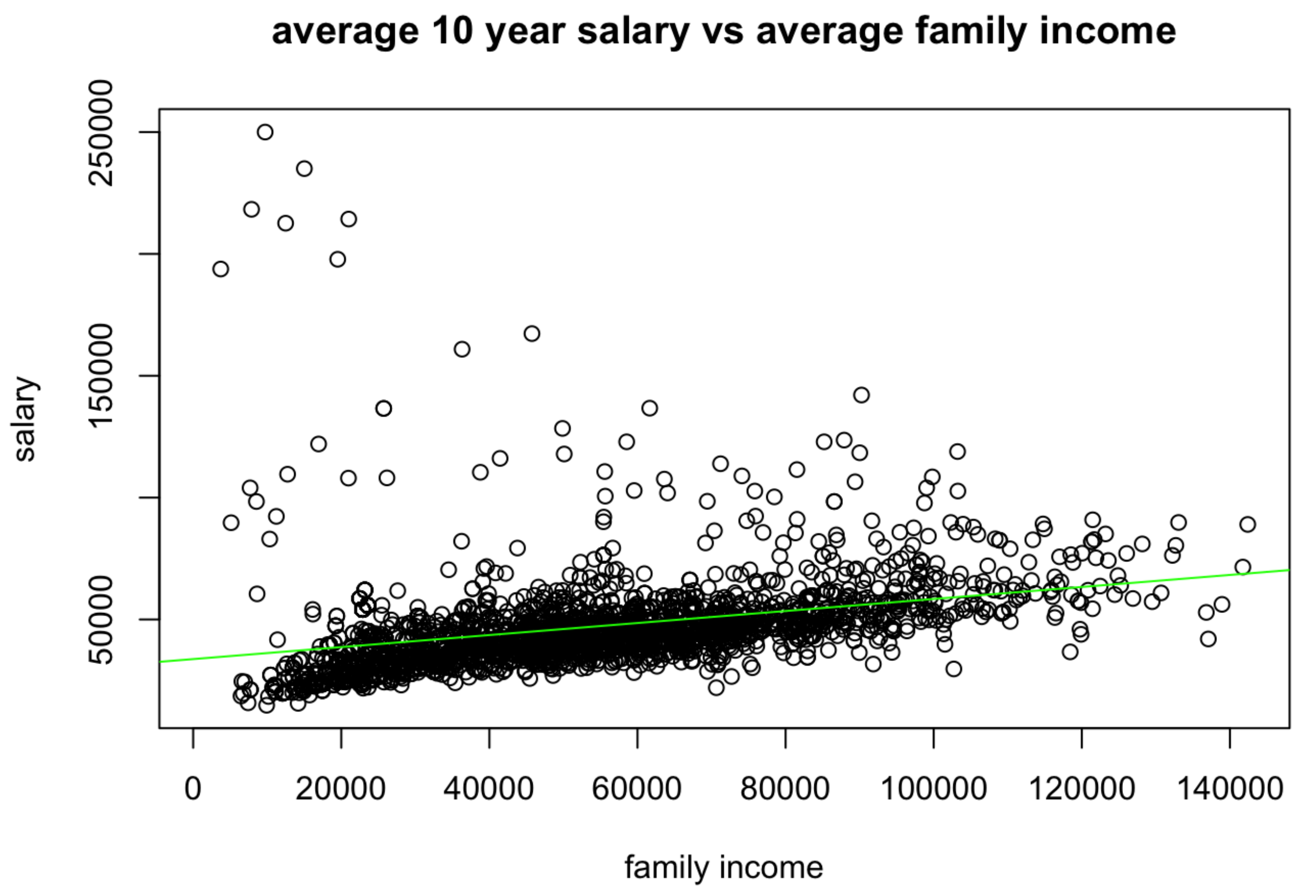
```
##
##                group1 group2
##   Less than 1   1106    125
##   More than 1    103     11
```

```
#barplot(prop.table(table(data9$group,data9$mt1b),margin = 2))
```

open admission is independent of group since we can see than both group 1 and 2 have no open admission.
Main campus is dependend on the group since we can see that only schools in group 1 doesn’t have main campus.
Ownership is independent on group since both groups have public and Nonprofit schools. However, For profit school only exists in group 1.
whether the school has more than 1 branch is independent on groups since both groups have less than 1 and more than 1 braches schools.

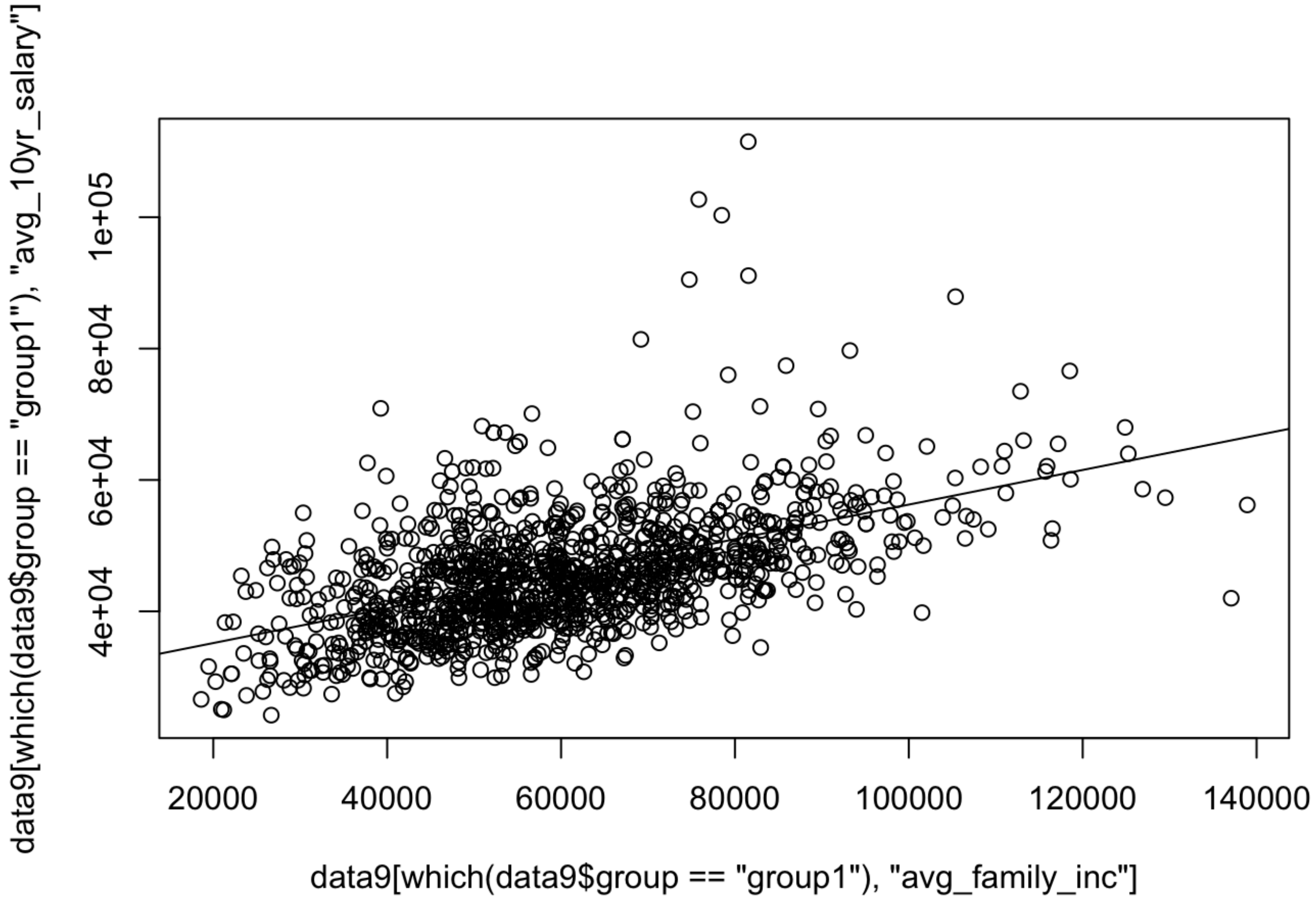
10. relationship between avg_10yr_salary using avg_family_inc for all schools. (a).

```
plot(data9$avg_family_inc,data9$avg_10yr_salary, main = "average 10 year salary vs
average family income", xlab = "family income",ylab = "salary")
model3 = lm(data9$avg_10yr_salary~data9$avg_family_inc)
abline(model3,col = "green")
```



Althogh there are lots of outliers in this graph, the regression model is linear. We can see that there is positive linear relationship between salary and family income. I suspect that it's group that affects the model. (b).

```
model4 = lm(data9[which(data9$group == "group1"),"avg_10yr_salary"]~data9[which(dat
a9$group == "group1"),"avg_family_inc"])
plot(data9[which(data9$group == "group1"),"avg_family_inc"],data9[which(data9$group
== "group1"),"avg_10yr_salary"])
abline(model4)
```



As we can see, if we use group1, we will get a more accurate regression line where not a lot of outliers are in the data.