Data Analytics: Assignment 1 (Group 4)

Load the data

```
if (!require("ISLR")) install.packages("ISLR")
```

```
## Loading required package: ISLR
```

```
library("ISLR")
data("College")
```

Task 1: Descriptive statistics and visualization

```
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
```

```
##  Max.   :118.00
```

```
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

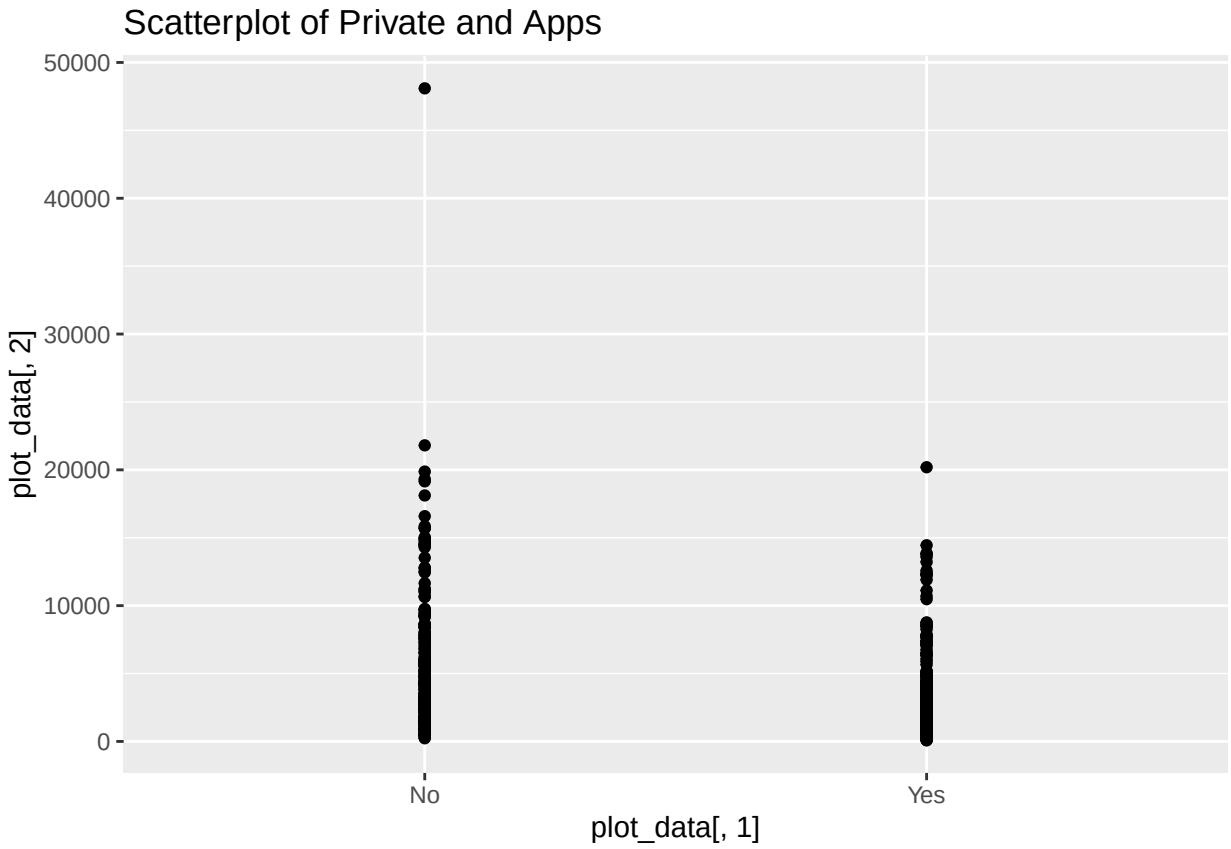
```
cor(College[, -1])
```

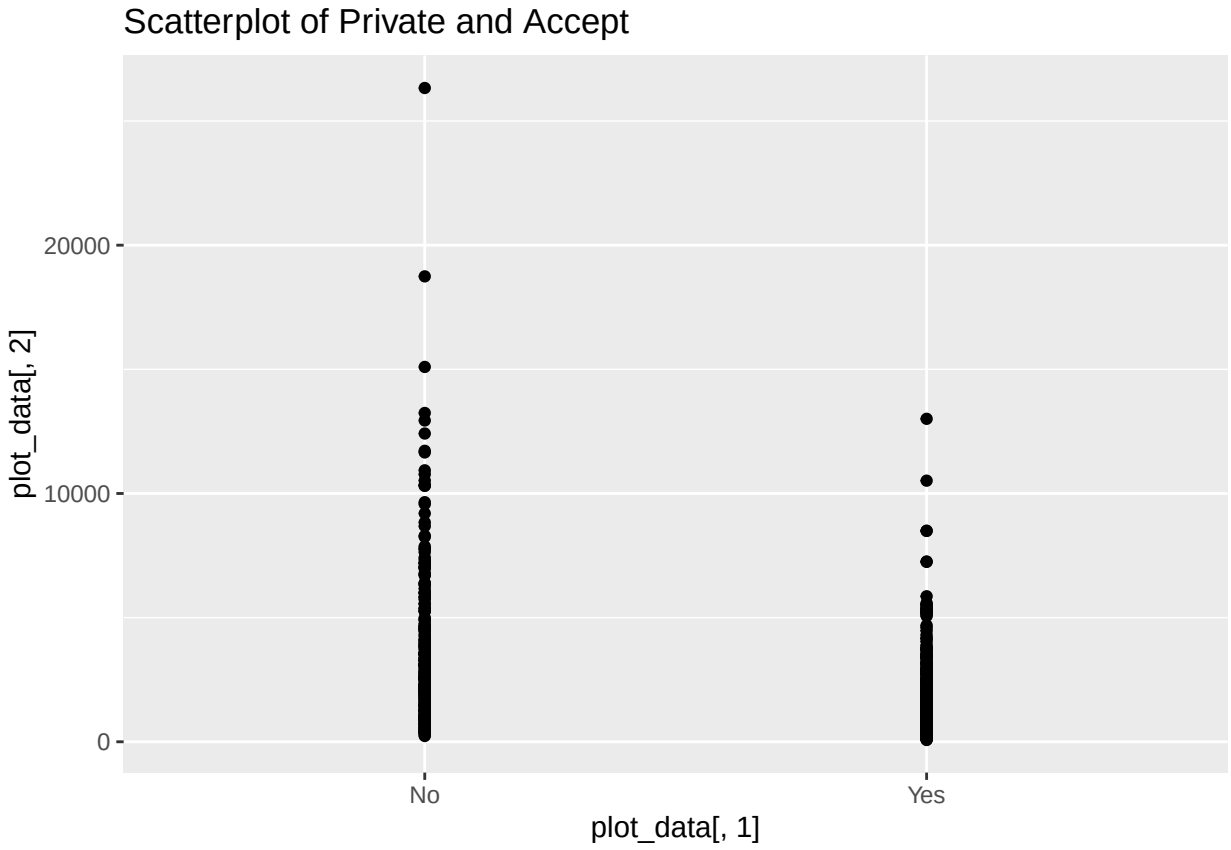
```
##                   Apps      Accept      Enroll  Top10perc   Top25perc
## Apps         1.00000000  0.94345057  0.84682205  0.3388337  0.35163990
## Accept       0.94345057  1.00000000  0.91163666  0.1924469  0.24747574
## Enroll       0.84682205  0.91163666  1.00000000  0.1812935  0.22674511
## Top10perc    0.33883368  0.19244693  0.18129353  1.0000000  0.89199497
## Top25perc    0.35163990  0.24747574  0.22674511  0.8919950  1.00000000
## F.Undergrad  0.81449058  0.87422328  0.96463965  0.1412887  0.19944466
## P.Undergrad  0.39826427  0.44127073  0.51306860 -0.1053563 -0.05357664
## Outstate     0.05015903 -0.02575455 -0.15547734  0.5623305  0.48939383
## Room.Board   0.16493896  0.09089863 -0.04023168  0.3714804  0.33148989
## Books        0.13255860  0.11352535  0.11271089  0.1188584  0.11552713
## Personal     0.17873085  0.20098867  0.28092946 -0.0933164 -0.08081027
## PhD          0.39069733  0.35575788  0.33146914  0.5318280  0.54586221
## Terminal     0.36949147  0.33758337  0.30827407  0.4911350  0.52474884
## S.F.Ratio    0.09563303  0.17622901  0.23727131 -0.3848745 -0.29462884
## perc.alumni -0.09022589 -0.15998987 -0.18079413  0.4554853  0.41786429
## Expend       0.25959198  0.12471701  0.06416923  0.6609134  0.52744743
## Grad.Rate    0.14675460  0.06731255 -0.02234104  0.4949892  0.47728116
##              F.Undergrad P.Undergrad    Outstate  Room.Board        Books
## Apps          0.81449058  0.39826427  0.05015903  0.16493896  0.132558598
## Accept        0.87422328  0.44127073 -0.02575455  0.09089863  0.113525352
## Enroll        0.96463965  0.51306860 -0.15547734 -0.04023168  0.112710891
## Top10perc     0.14128873 -0.10535628  0.56233054  0.37148038  0.118858431
## Top25perc     0.19944466 -0.05357664  0.48939383  0.33148989  0.115527130
## F.Undergrad   1.00000000  0.57051219 -0.21574200 -0.06889039  0.115549761
## P.Undergrad   0.57051219  1.00000000 -0.25351232 -0.06132551  0.081199521
## Outstate     -0.21574200 -0.25351232  1.00000000  0.65425640  0.038854868
## Room.Board   -0.06889039 -0.06132551  0.65425640  1.00000000  0.127962970
## Books         0.11554976  0.08119952  0.03885487  0.12796297  1.000000000
```

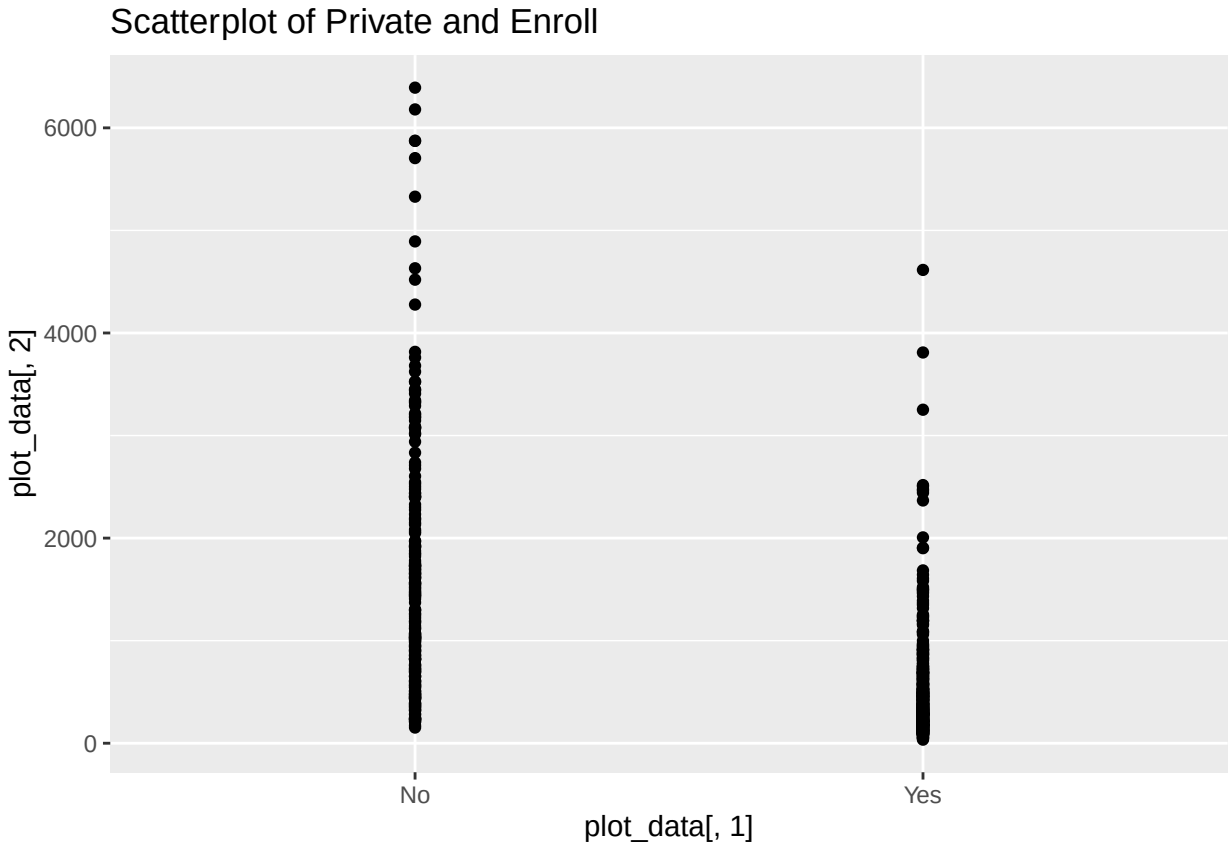
```
## Personal     0.31719954  0.31988162 -0.29908690 -0.19942818  0.179294764
## PhD          0.31833697  0.14911422  0.38298241  0.32920228  0.026905731
## Terminal     0.30001894  0.14190357  0.40798320  0.37453955  0.099954700
## S.F.Ratio    0.27970335  0.23253051 -0.55482128 -0.36262774 -0.031929274
## perc.alumni -0.22946222 -0.28079236  0.56626242  0.27236345 -0.040207736
## Expend       0.01865162 -0.08356842  0.67277862  0.50173942  0.112409075
## Grad.Rate   -0.07877313 -0.25700099  0.57128993  0.42494154  0.001060894
##                Personal         PhD    Terminal   S.F.Ratio perc.alumni
## Apps         0.17873085  0.39069733  0.36949147  0.09563303 -0.09022589
## Accept       0.20098867  0.35575788  0.33758337  0.17622901 -0.15998987
## Enroll       0.28092946  0.33146914  0.30827407  0.23727131 -0.18079413
## Top10perc   -0.09331640  0.53182802  0.49113502 -0.38487451  0.45548526
## Top25perc   -0.08081027  0.54586221  0.52474884 -0.29462884  0.41786429
## F.Undergrad  0.31719954  0.31833697  0.30001894  0.27970335 -0.22946222
## P.Undergrad  0.31988162  0.14911422  0.14190357  0.23253051 -0.28079236
## Outstate    -0.29908690  0.38298241  0.40798320 -0.55482128  0.56626242
## Room.Board  -0.19942818  0.32920228  0.37453955 -0.36262774  0.27236345
## Books        0.17929476  0.02690573  0.09995470 -0.03192927 -0.04020774
## Personal     1.00000000 -0.01093579 -0.03061311  0.13634483 -0.28596808
## PhD         -0.01093579  1.00000000  0.84958703 -0.13053011  0.24900866
## Terminal    -0.03061311  0.84958703  1.00000000 -0.16010395  0.26713029
## S.F.Ratio    0.13634483 -0.13053011 -0.16010395  1.00000000 -0.40292917
## perc.alumni -0.28596808  0.24900866  0.26713029 -0.40292917  1.00000000
## Expend      -0.09789189  0.43276168  0.43879922 -0.58383204  0.41771172
## Grad.Rate   -0.26934396  0.30503785  0.28952723 -0.30671041  0.49089756
##                 Expend    Grad.Rate
## Apps         0.25959198  0.146754600
## Accept       0.12471701  0.067312550
## Enroll       0.06416923 -0.022341039
## Top10perc    0.66091341  0.494989235
## Top25perc    0.52744743  0.477281164
## F.Undergrad  0.01865162 -0.078773129
## P.Undergrad -0.08356842 -0.257000991
## Outstate     0.67277862  0.571289928
## Room.Board   0.50173942  0.424941541
## Books        0.11240908  0.001060894
## Personal    -0.09789189 -0.269343964
## PhD          0.43276168  0.305037850
## Terminal     0.43879922  0.289527232
## S.F.Ratio   -0.58383204 -0.306710405
## perc.alumni  0.41771172  0.490897562
## Expend       1.00000000  0.390342696
## Grad.Rate    0.39034270  1.000000000
```

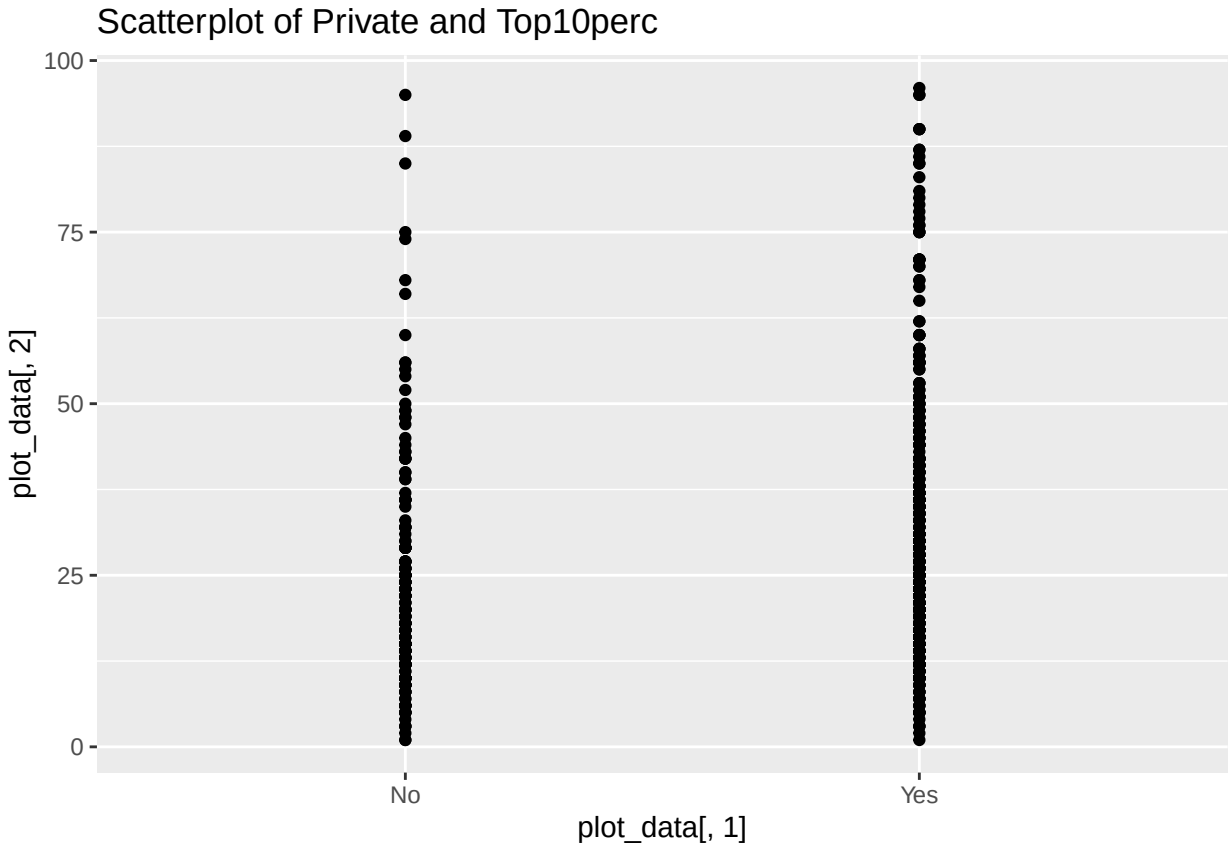
```r
library(ggplot2)

# Subset of variables to plot
cols <- c("Private", "Apps", "Accept", "Enroll", "Top10perc", "Top25perc", "F.Undergrad", "P.Undergrad"

# Create scatterplots for all pairs of variables
for (i in 1:(length(cols) - 1)) {
  for (j in (i+1):length(cols)) {
    plot_data <- College[, c(cols[i], cols[j])]
    plot <- ggplot(plot_data, aes(x = plot_data[,1], y = plot_data[,2])) +
```

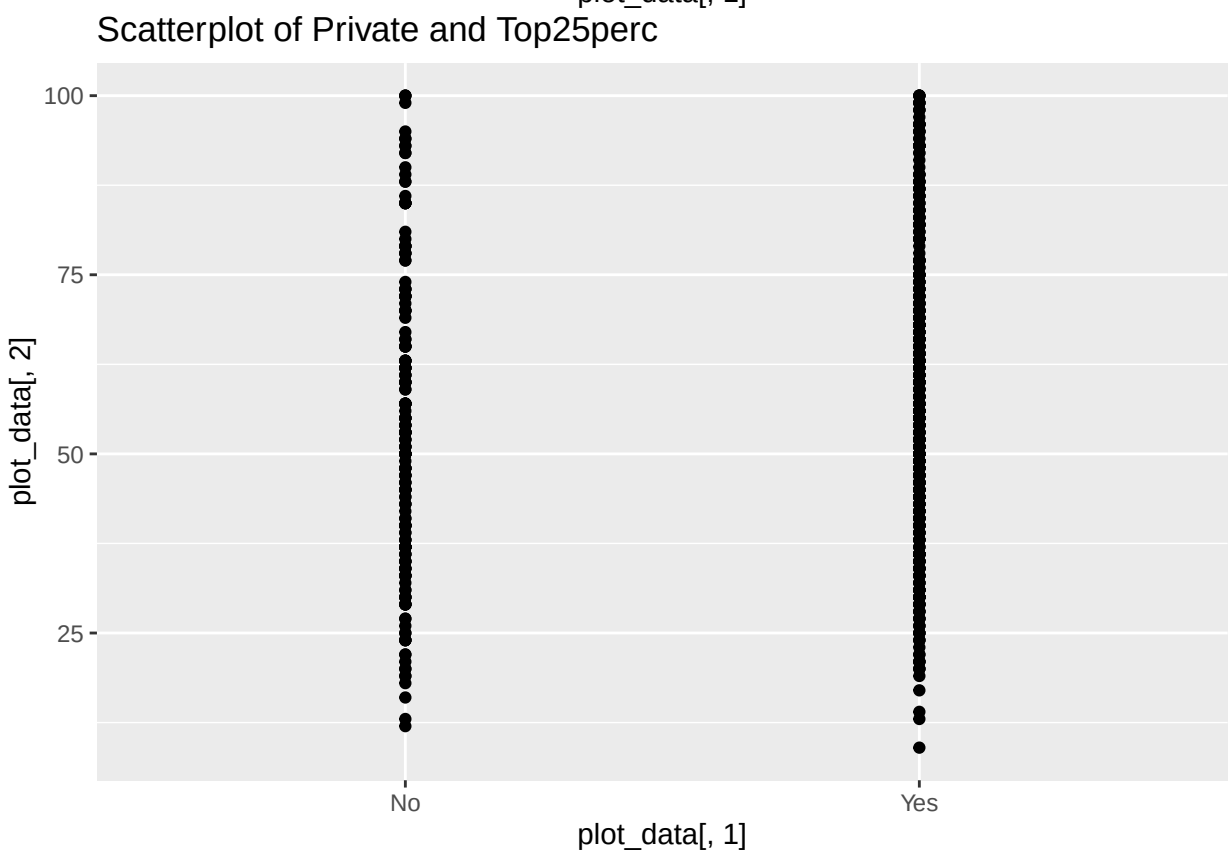
```
      geom_point() +
      ggtitle(paste("Scatterplot of", cols[i], "and", cols[j]))
    print(plot)
  }
}
```

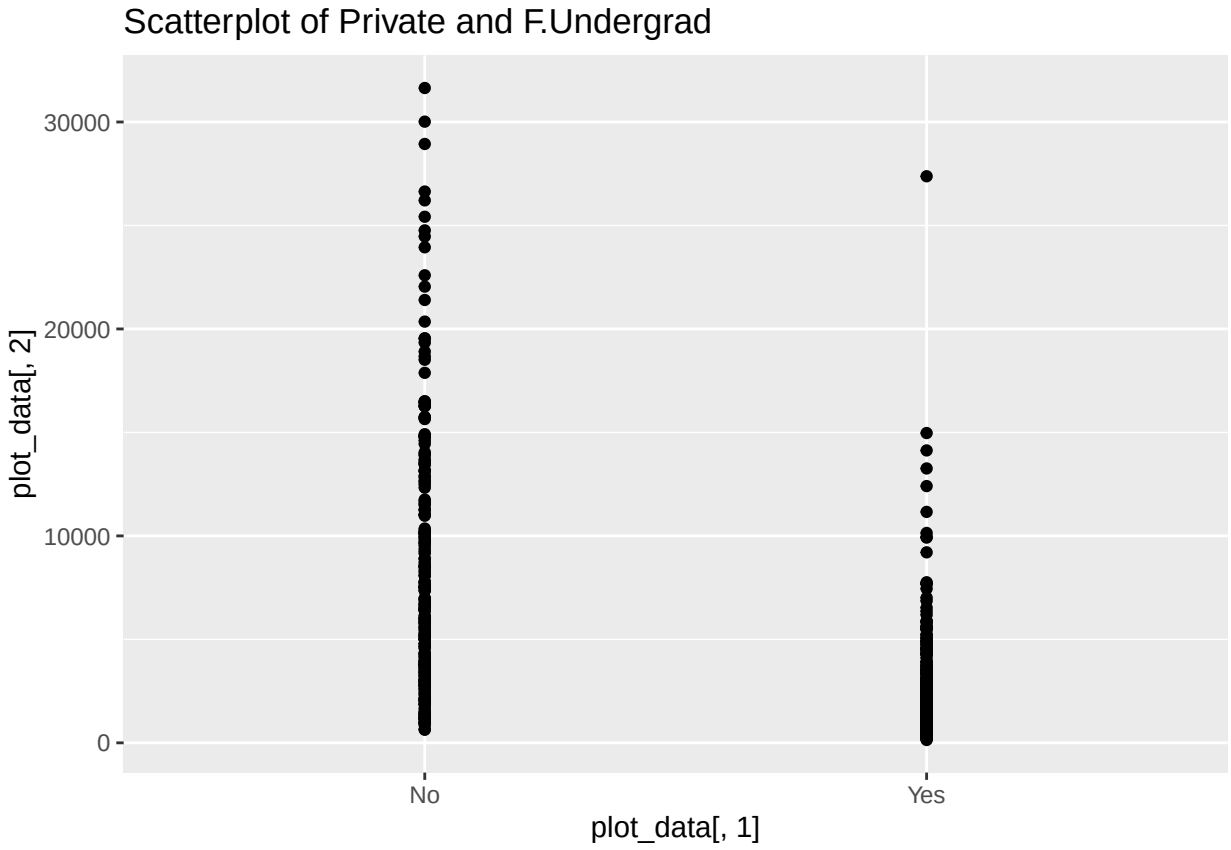


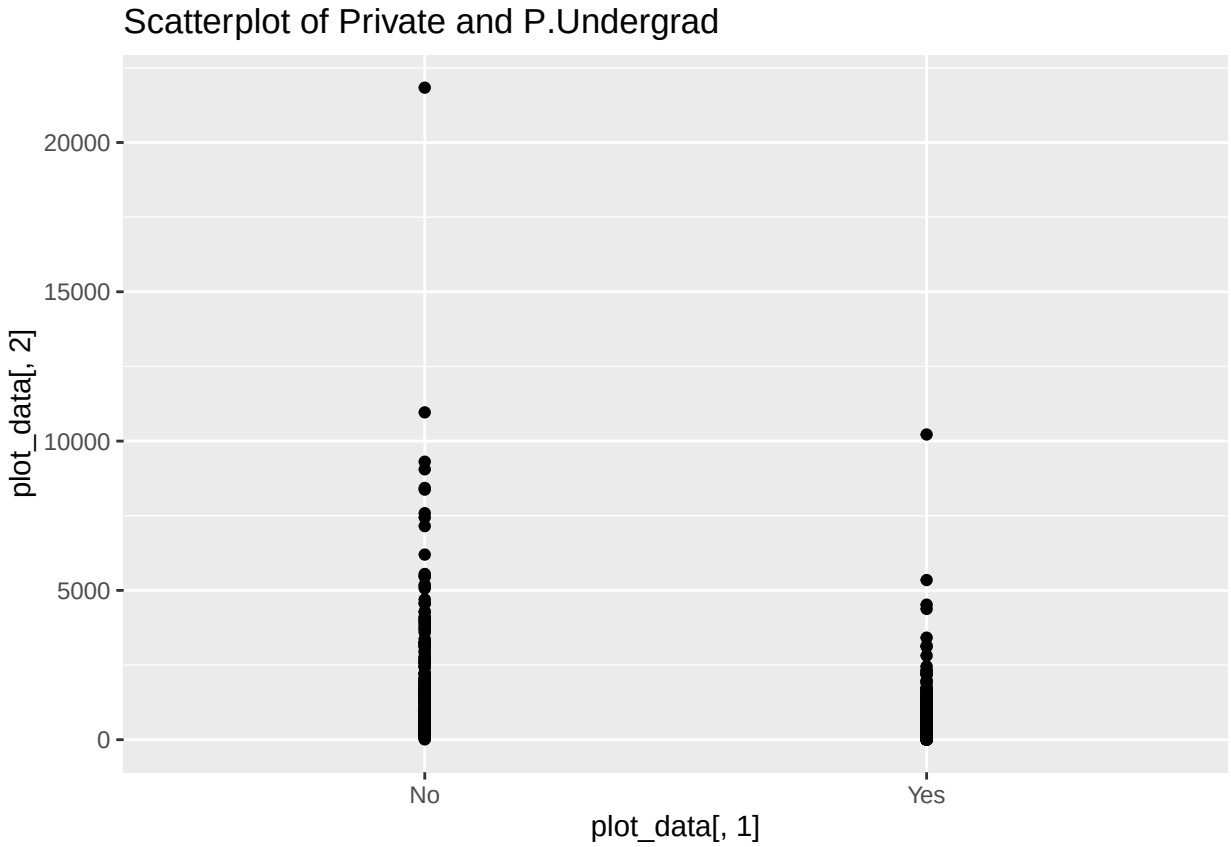

Scatterplot of Private and Accept
20000
10000
0
plot_data[, 2]
No
Yes
plot_data[, 1]

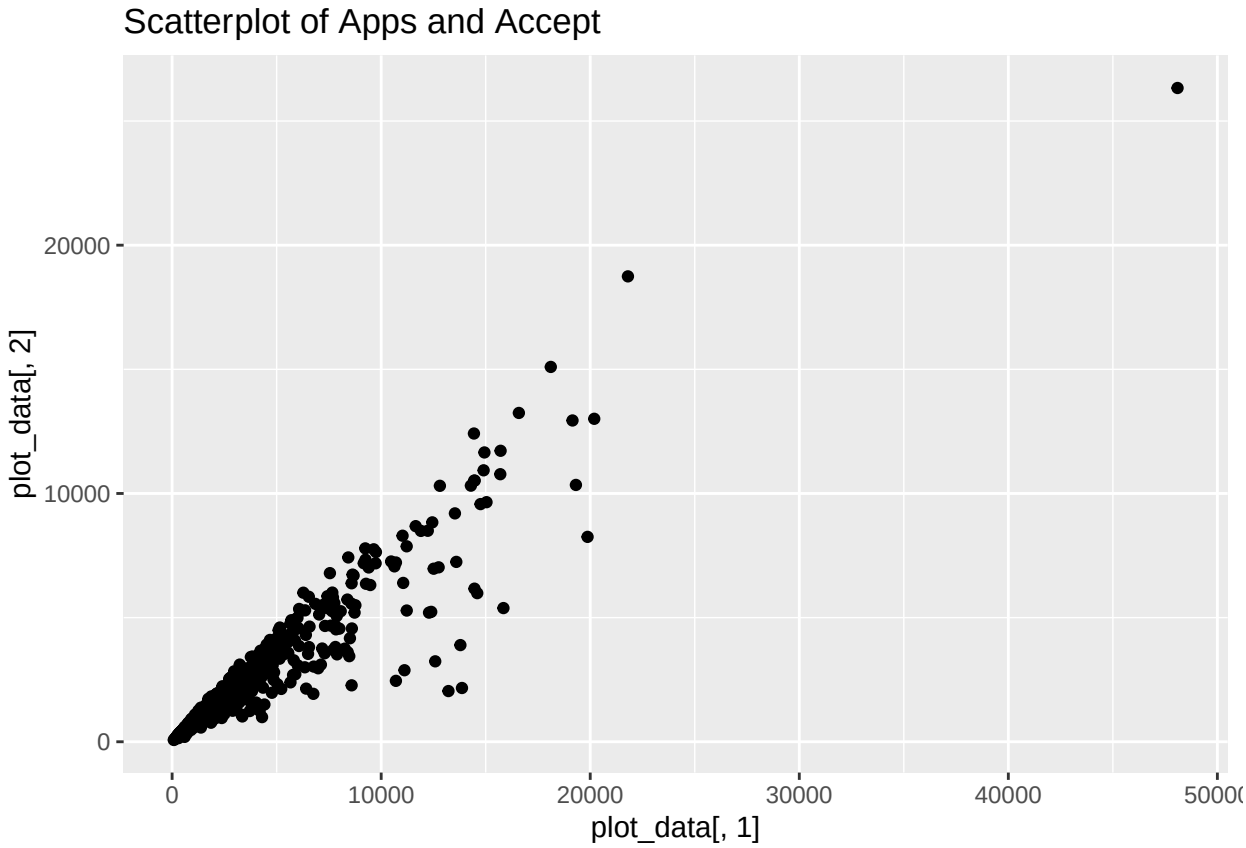

Scatterplot of Private and Enroll
6000
4000
2000
0
plot_data[, 2]
No
Yes
plot_data[, 1]

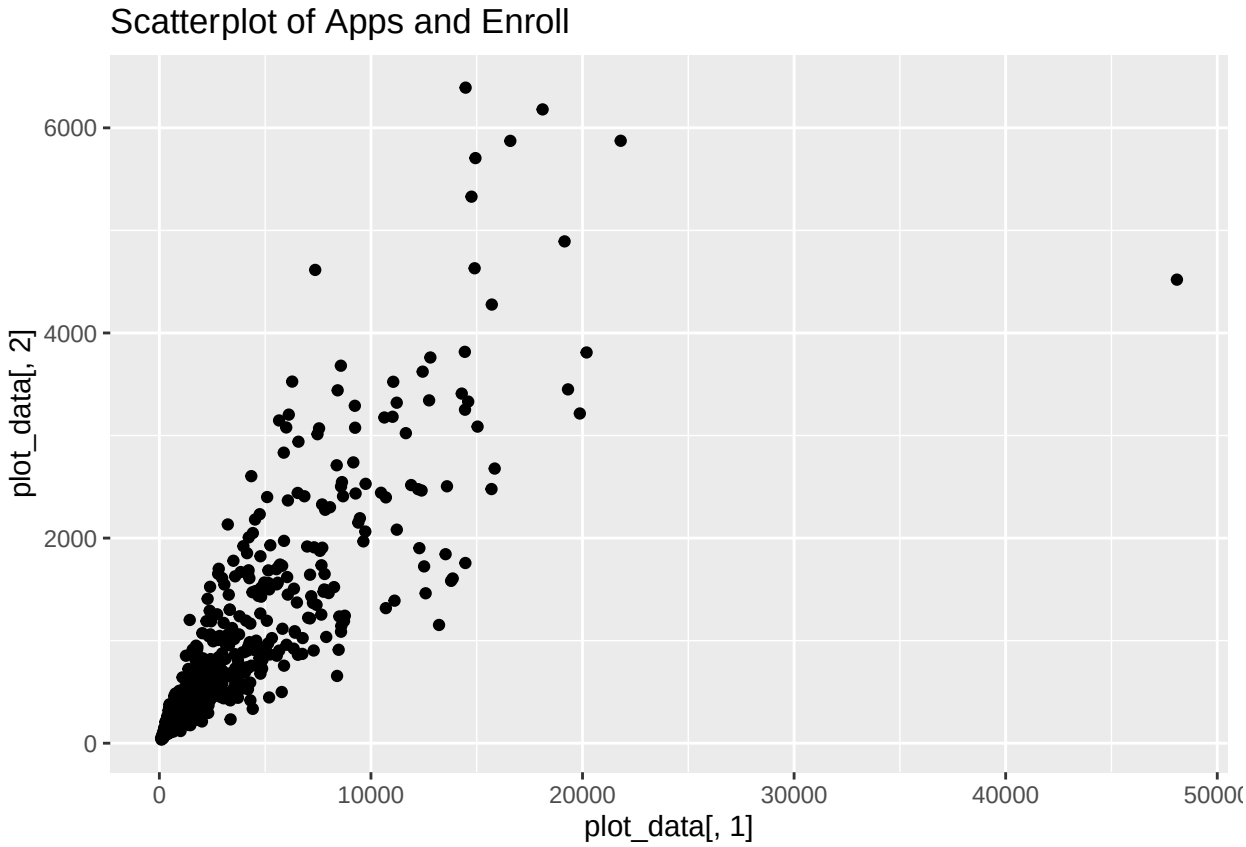

Scatterplot of Private and Top10perc
100
75
50
25
0
plot_data[, 2]
No
Yes
plot_data[, 1]

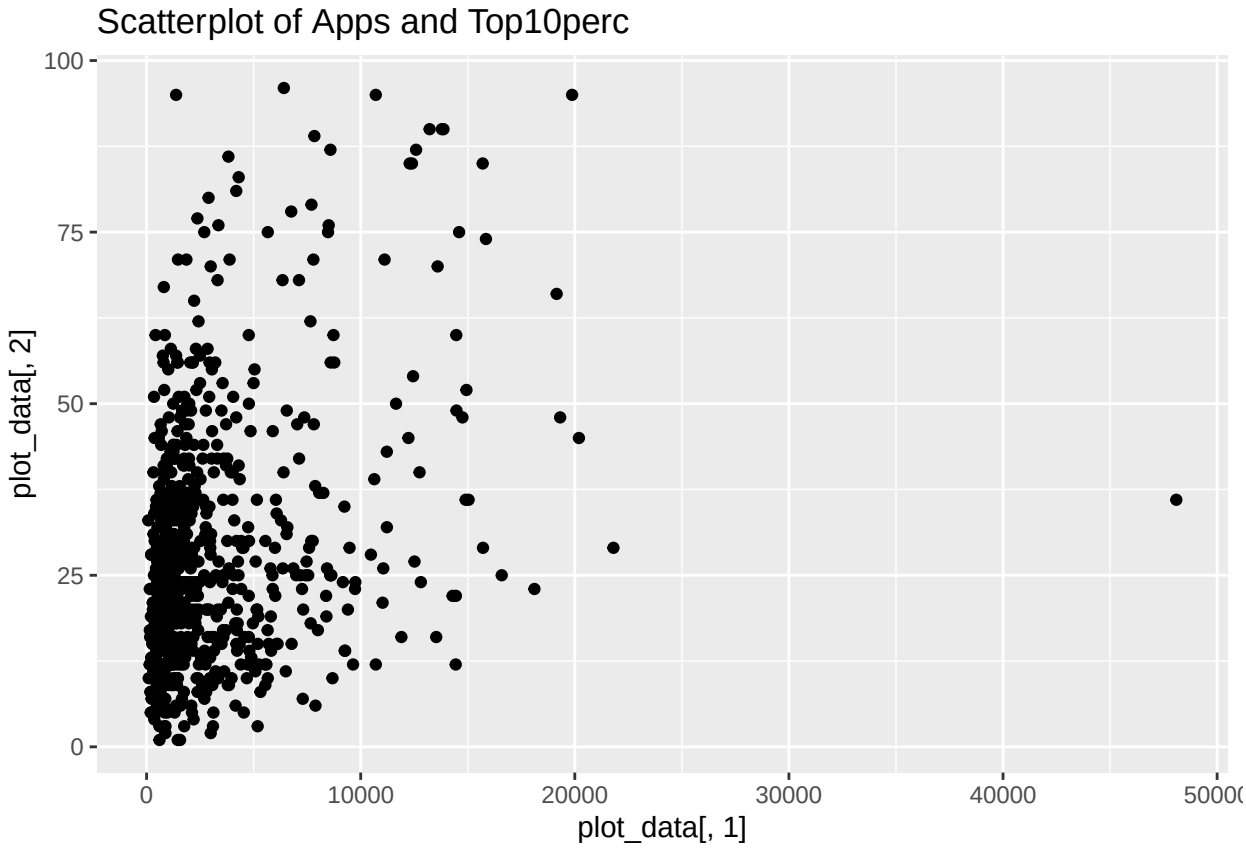

Scatterplot of Private and Top25perc
100
75
50
25
plot_data[, 2]
No
Yes
plot_data[, 1]

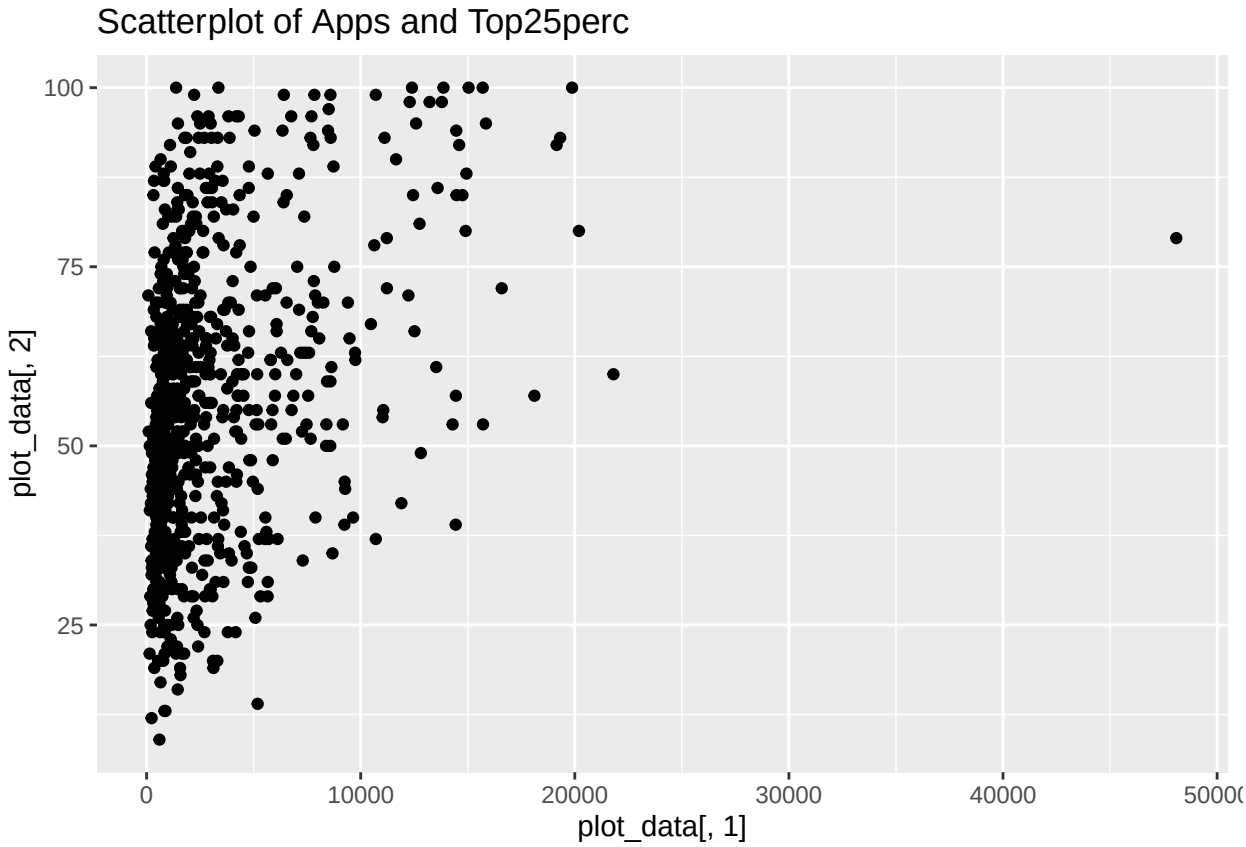

Scatterplot of Private and F.Undergrad
30000
20000
10000
0
plot_data[, 2]
No
Yes
plot_data[, 1]

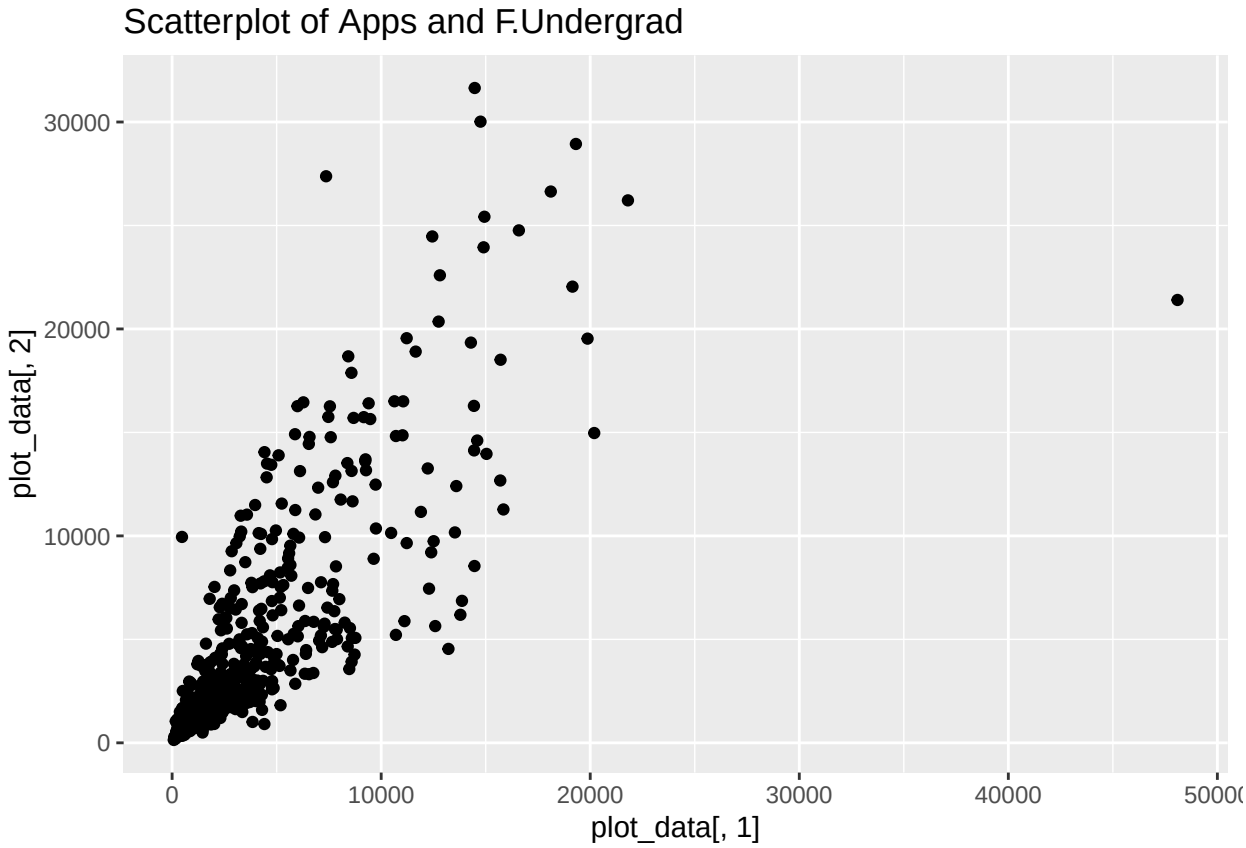

Scatterplot of Private and P.Undergrad
20000
15000
10000
5000
0
plot_data[, 2]
No
Yes
plot_data[, 1]

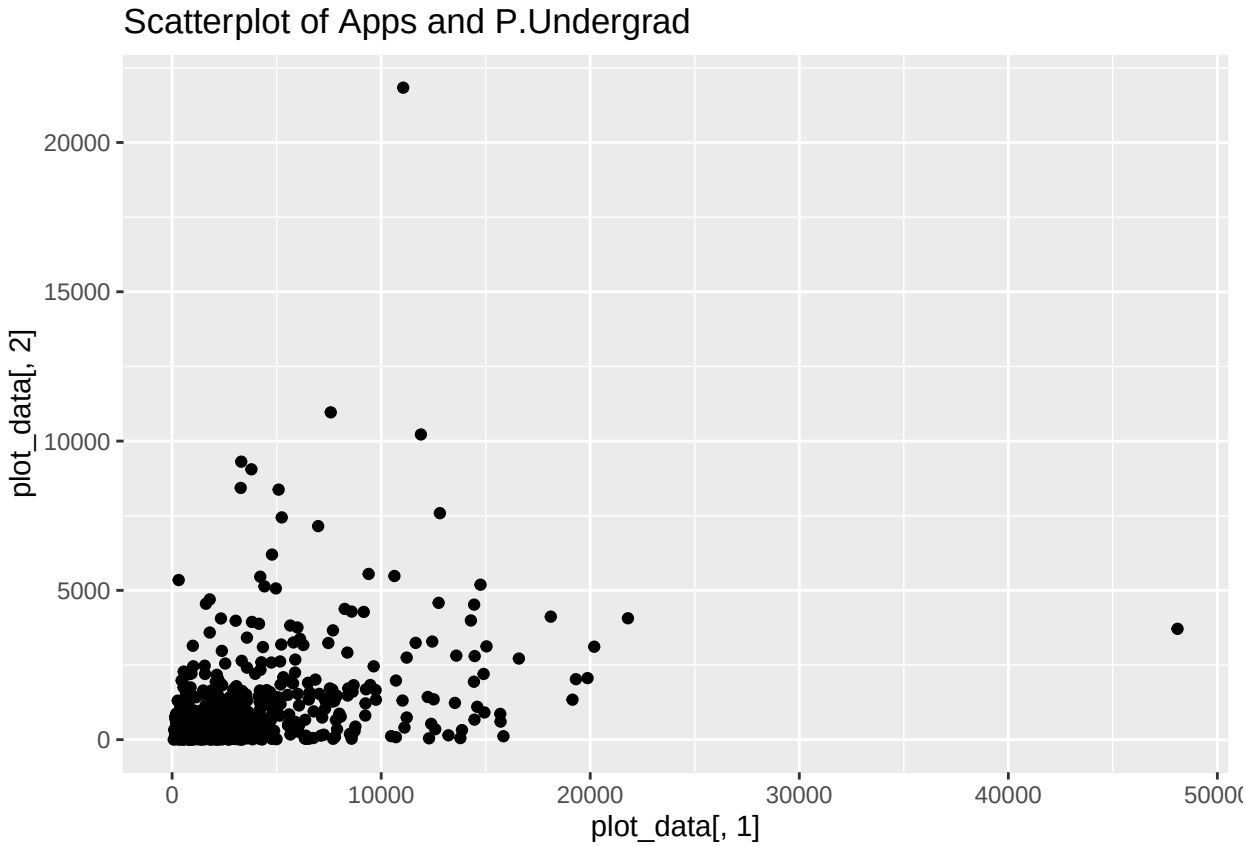

Scatterplot of Apps and Accept
plot_data[, 2]
0
10000
20000
0
10000
20000
30000
40000
plot_data[, 1]

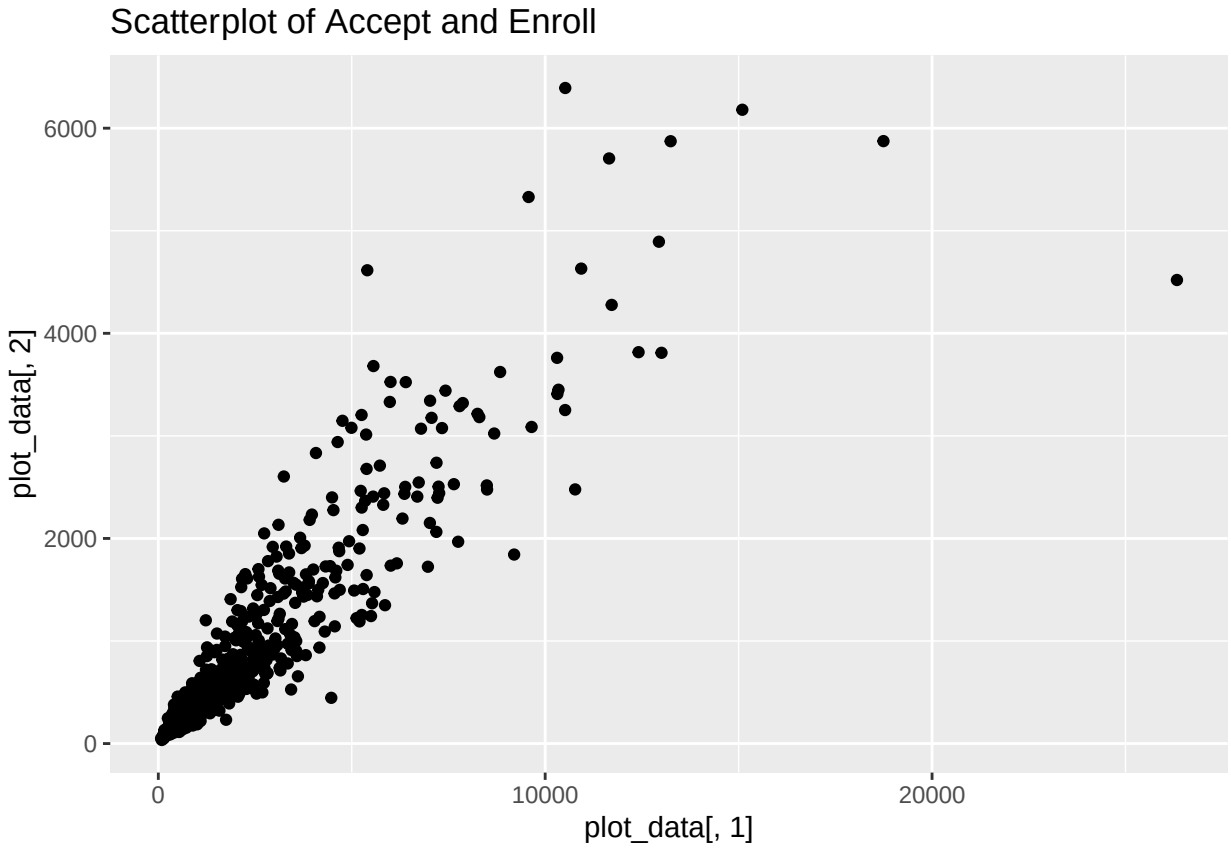

Scatterplot of Apps and Enroll
plot_data[, 2]
0
2000
4000
6000
0
10000
20000
30000
40000
plot_data[, 1]

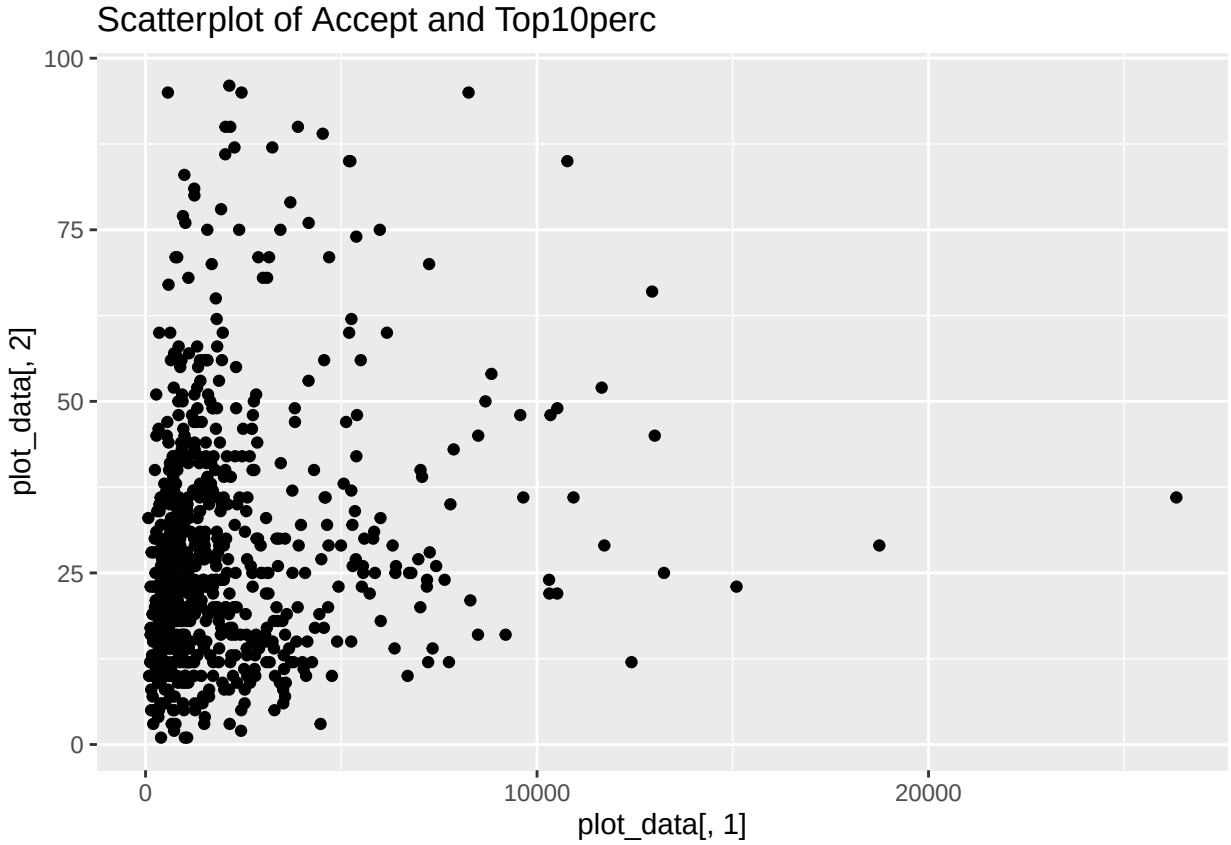

Scatterplot of Apps and Top10perc
100
75
50
25
0
plot_data[, 2]
0
10000
20000
30000
40000
plot_data[, 1]

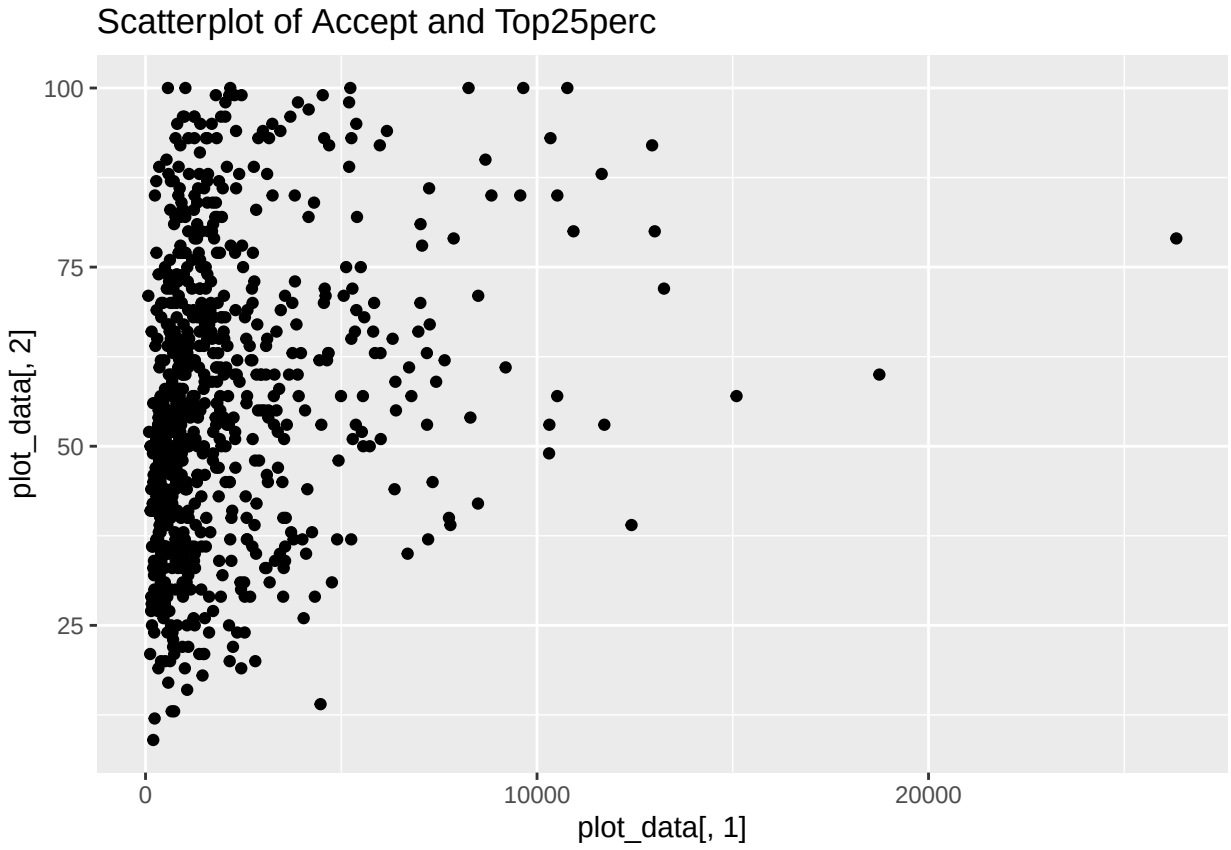

Scatterplot of Apps and Top25perc
100
75
50
25
plot_data[, 2]
0
10000
20000
30000
40000
plot_data[, 1]

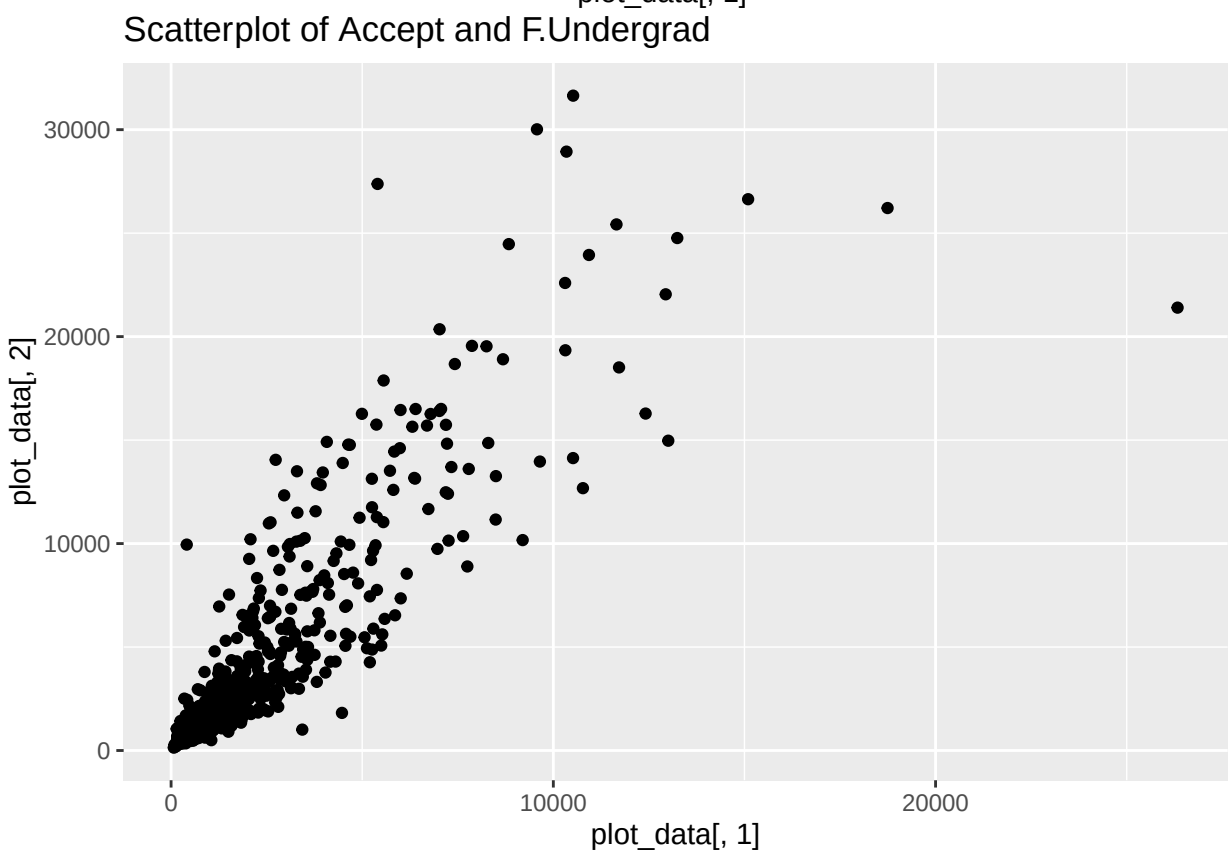

Scatterplot of Apps and F.Undergrad
plot_data[, 2]
30000
20000
10000
0
0
10000
20000
30000
40000
plot_data[, 1]

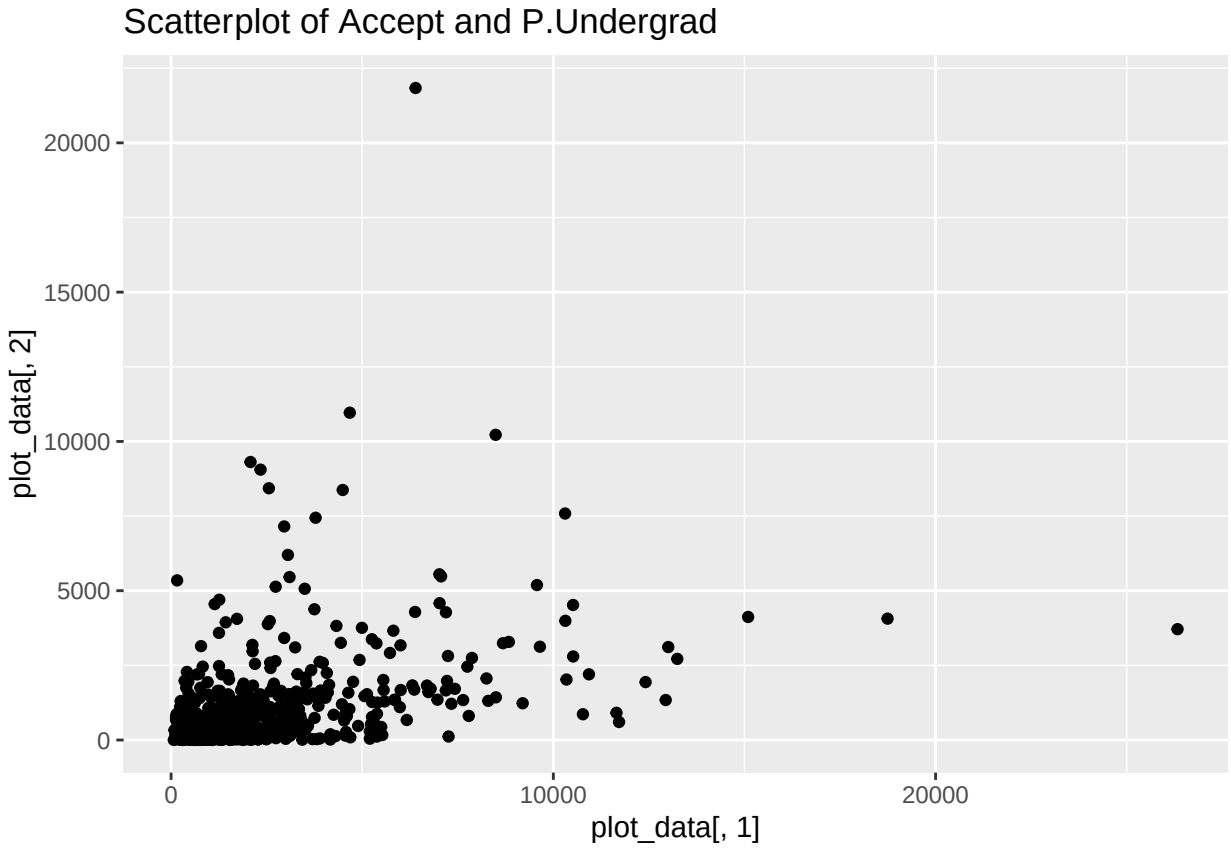

Scatterplot of Apps and P.Undergrad
plot_data[, 2]
20000
15000
10000
5000
0
0
10000
20000
30000
40000
plot_data[, 1]

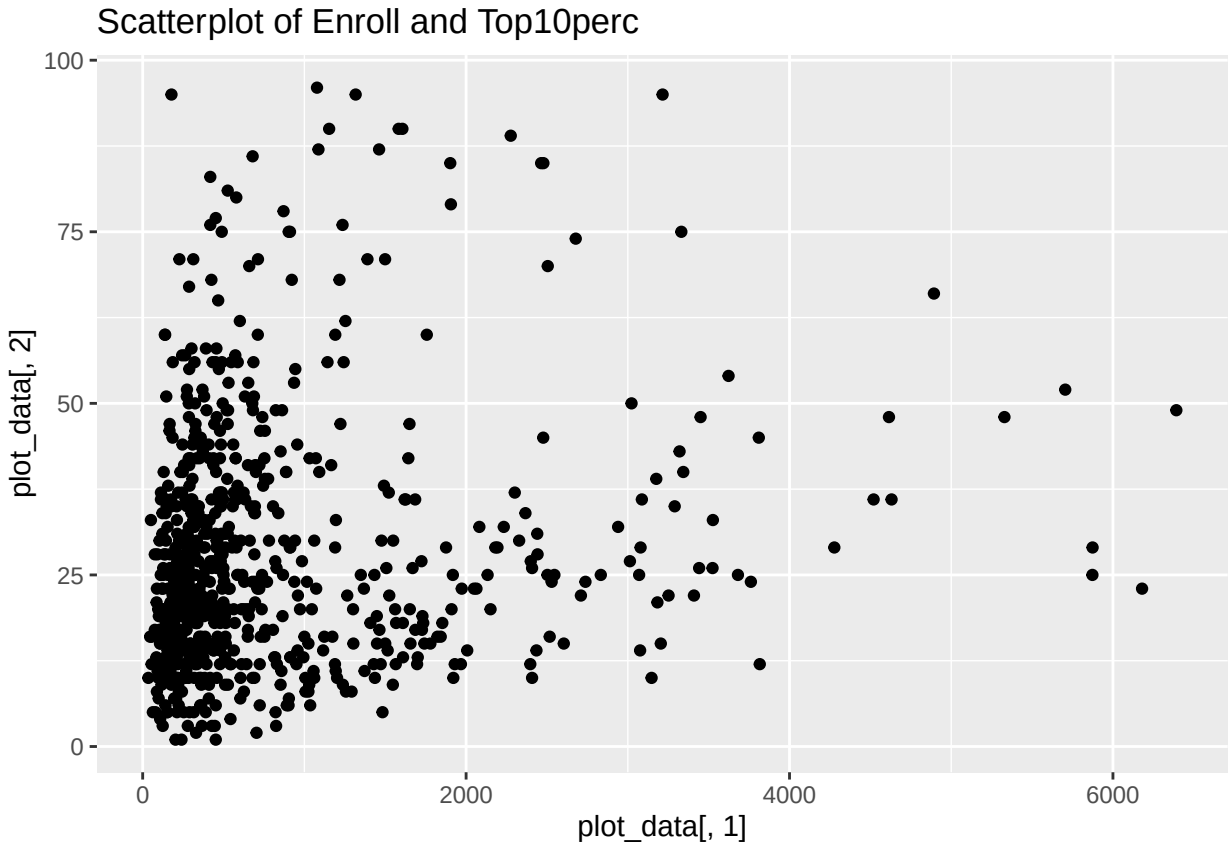

Scatterplot of Accept and Enroll
6000
4000
2000
0
plot_data[, 2]
0
10000
20000
plot_data[, 1]

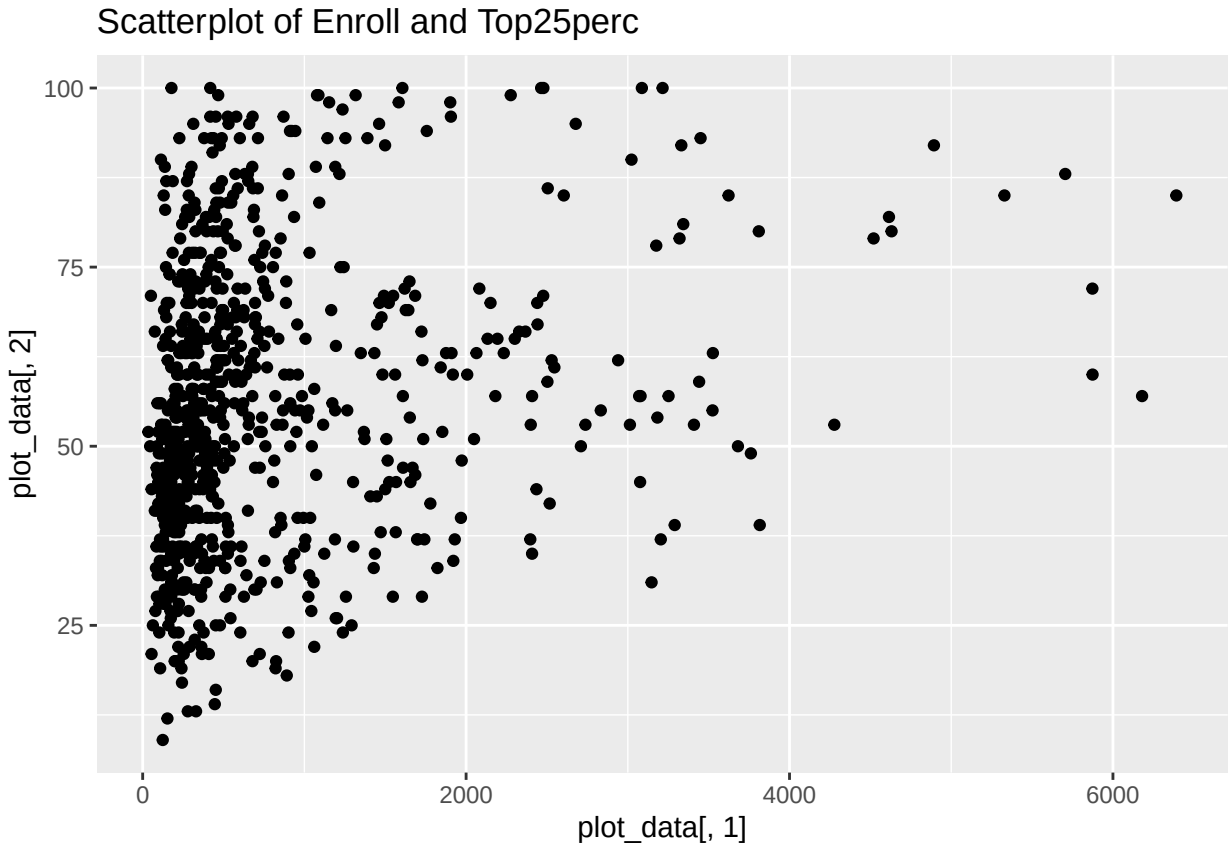

Scatterplot of Accept and Top10perc
100
75
50
25
0
plot_data[, 2]
0
10000
20000
plot_data[, 1]

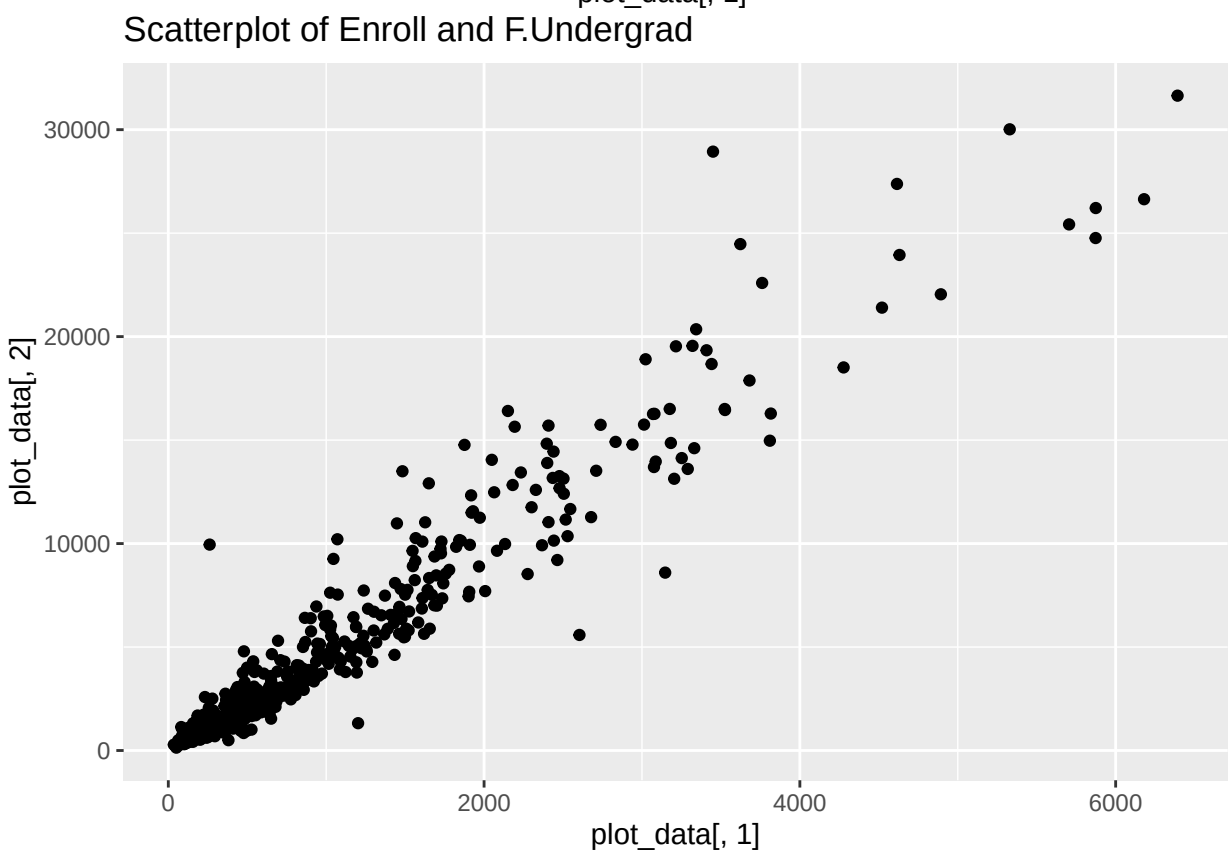

Scatterplot of Accept and Top25perc
100
75
50
25
plot_data[, 2]
0
10000
20000
plot_data[, 1]

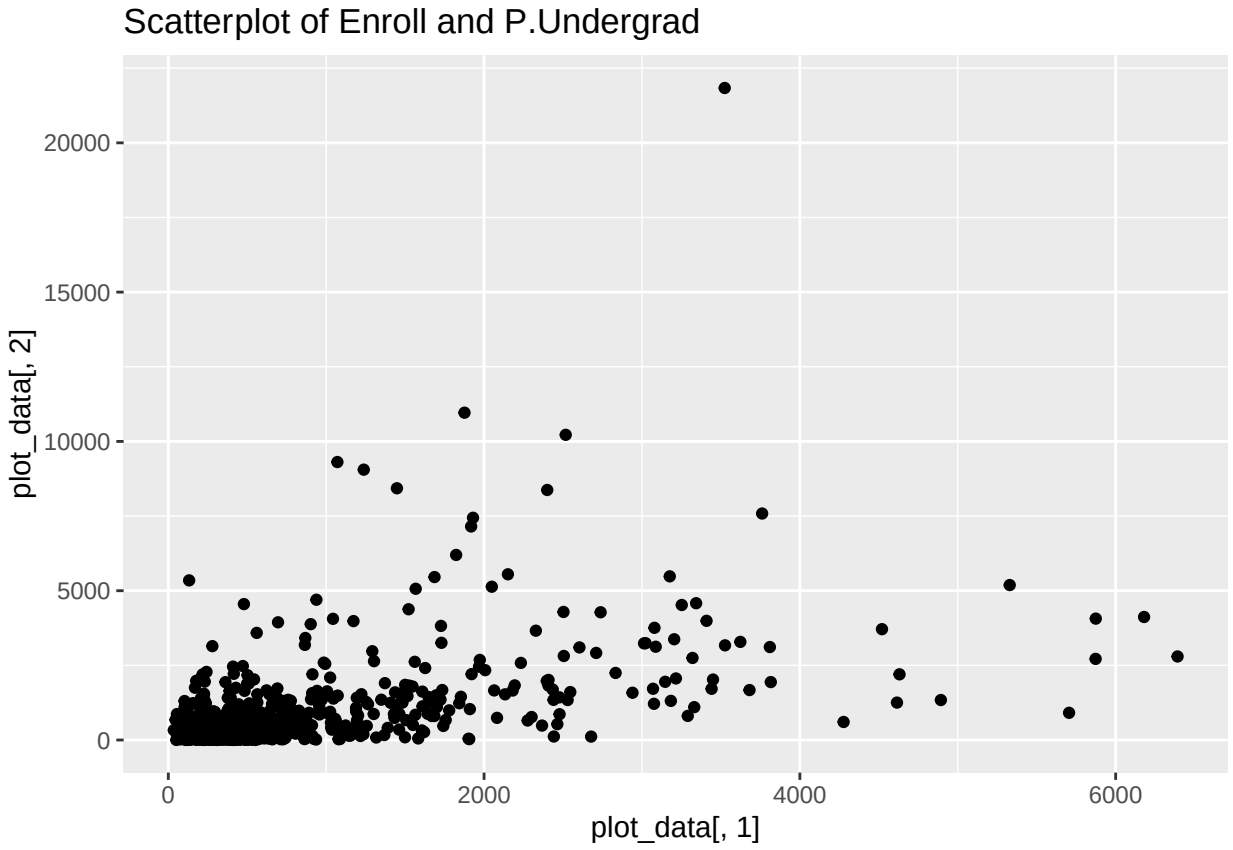

Scatterplot of Accept and F.Undergrad
30000
20000
10000
0
plot_data[, 2]
0
10000
20000
plot_data[, 1]

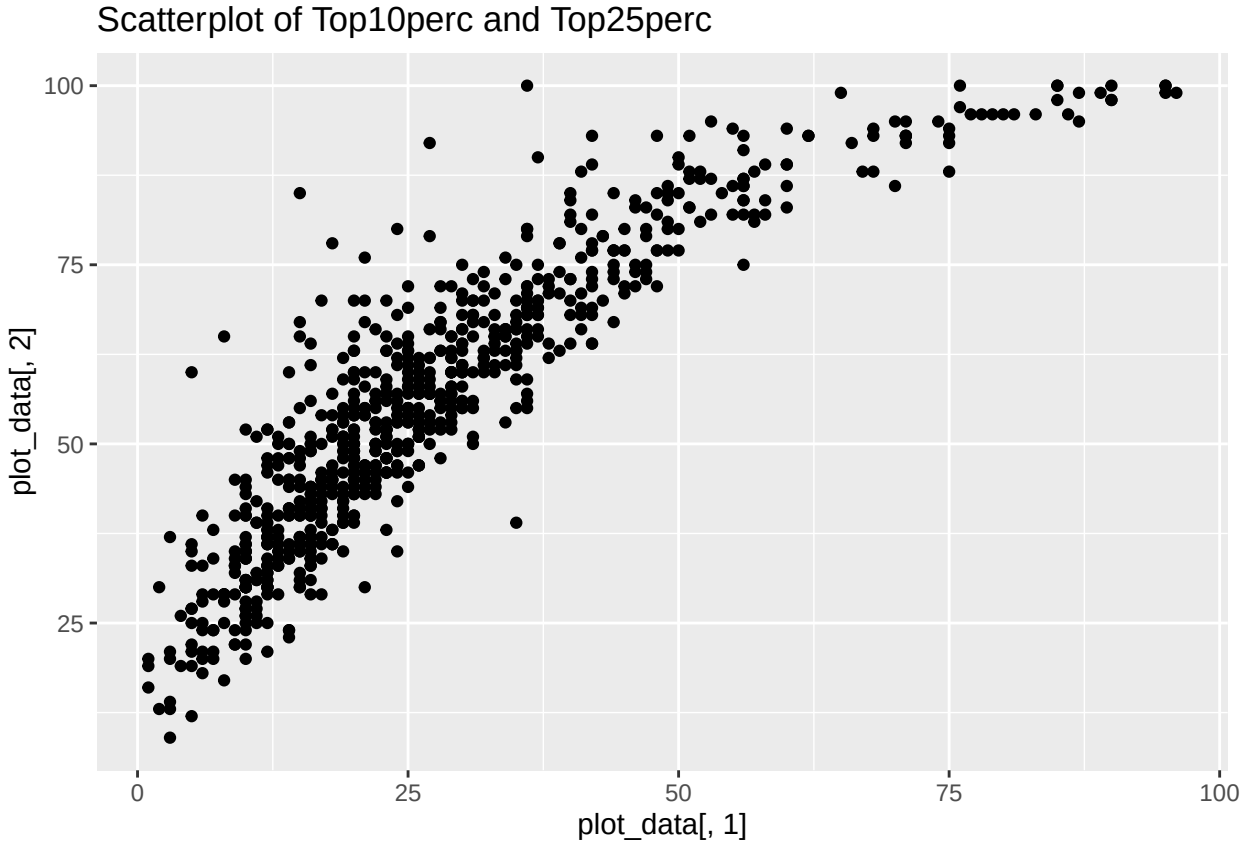

Scatterplot of Accept and P.Undergrad
20000
15000
10000
5000
0
plot_data[, 2]
0
10000
20000
plot_data[, 1]

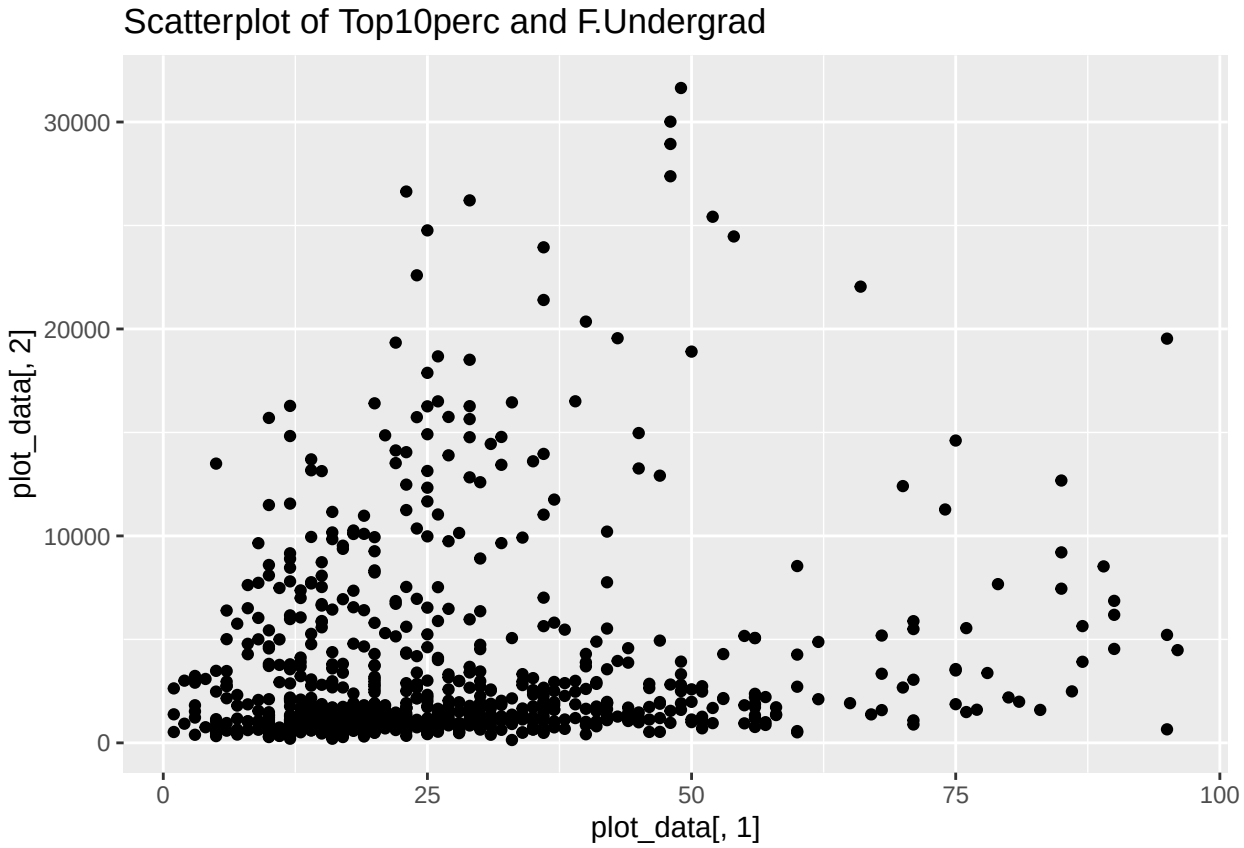

Scatterplot of Enroll and Top10perc
100
75
50
25
0
plot_data[, 2]
0
2000
4000
6000
plot_data[, 1]

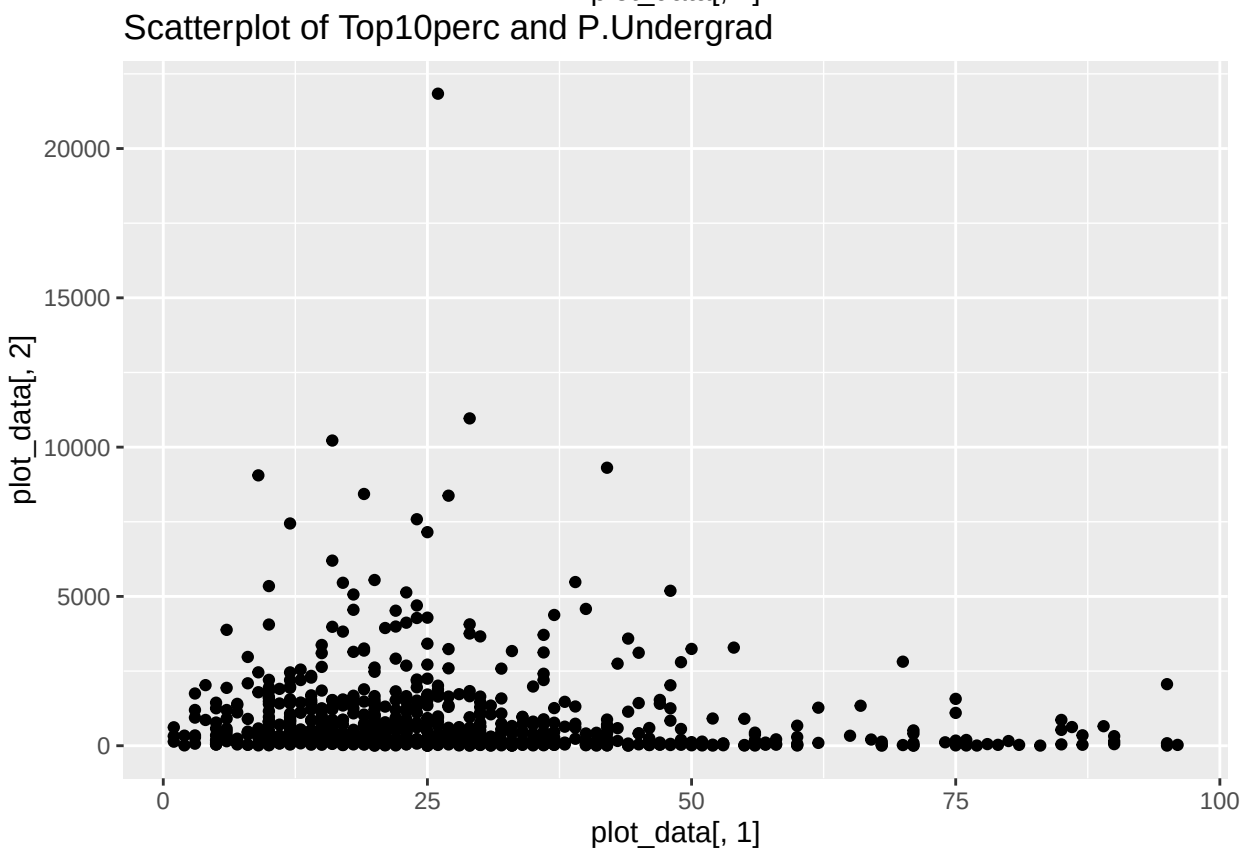

Scatterplot of Enroll and Top25perc
100
75
50
25
plot_data[, 2]
0
2000
4000
6000
plot_data[, 1]

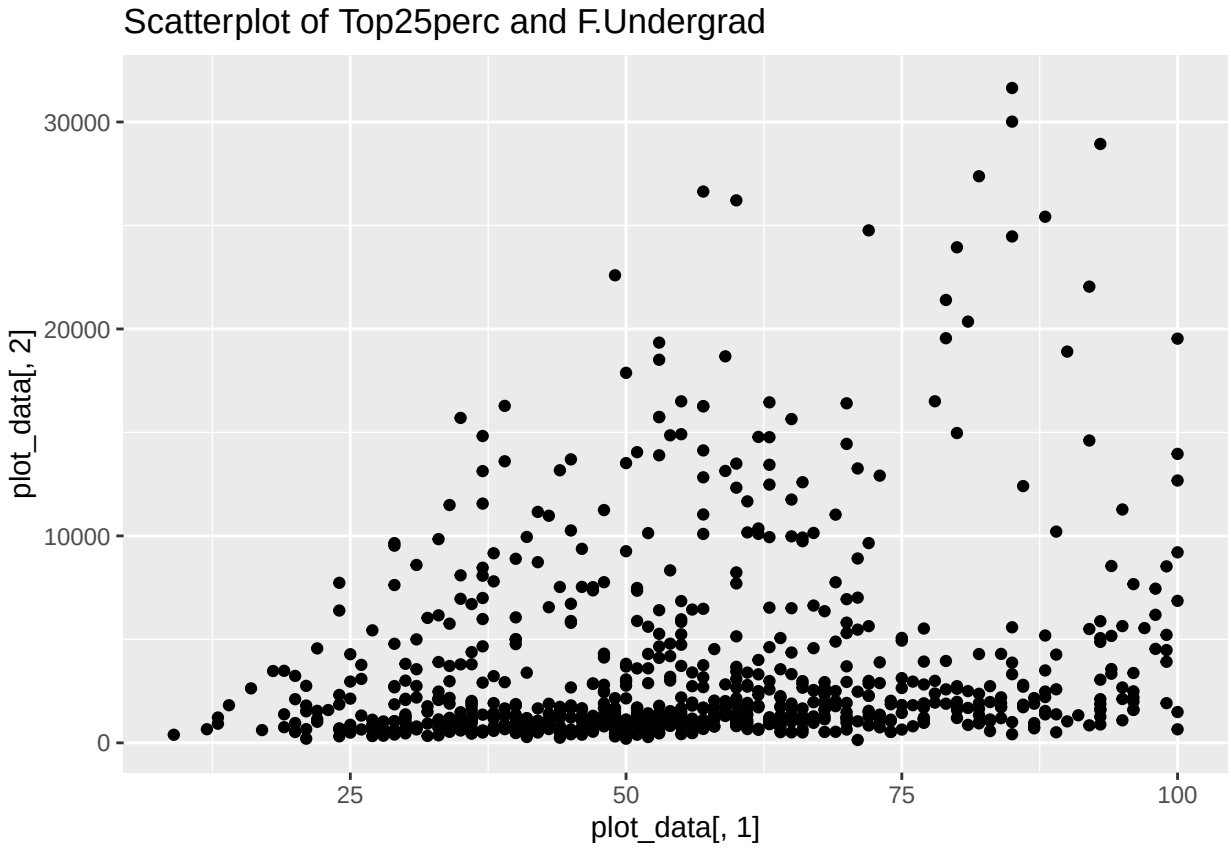

Scatterplot of Enroll and F.Undergrad
30000
20000
10000
0
plot_data[, 2]
0
2000
4000
6000
plot_data[, 1]

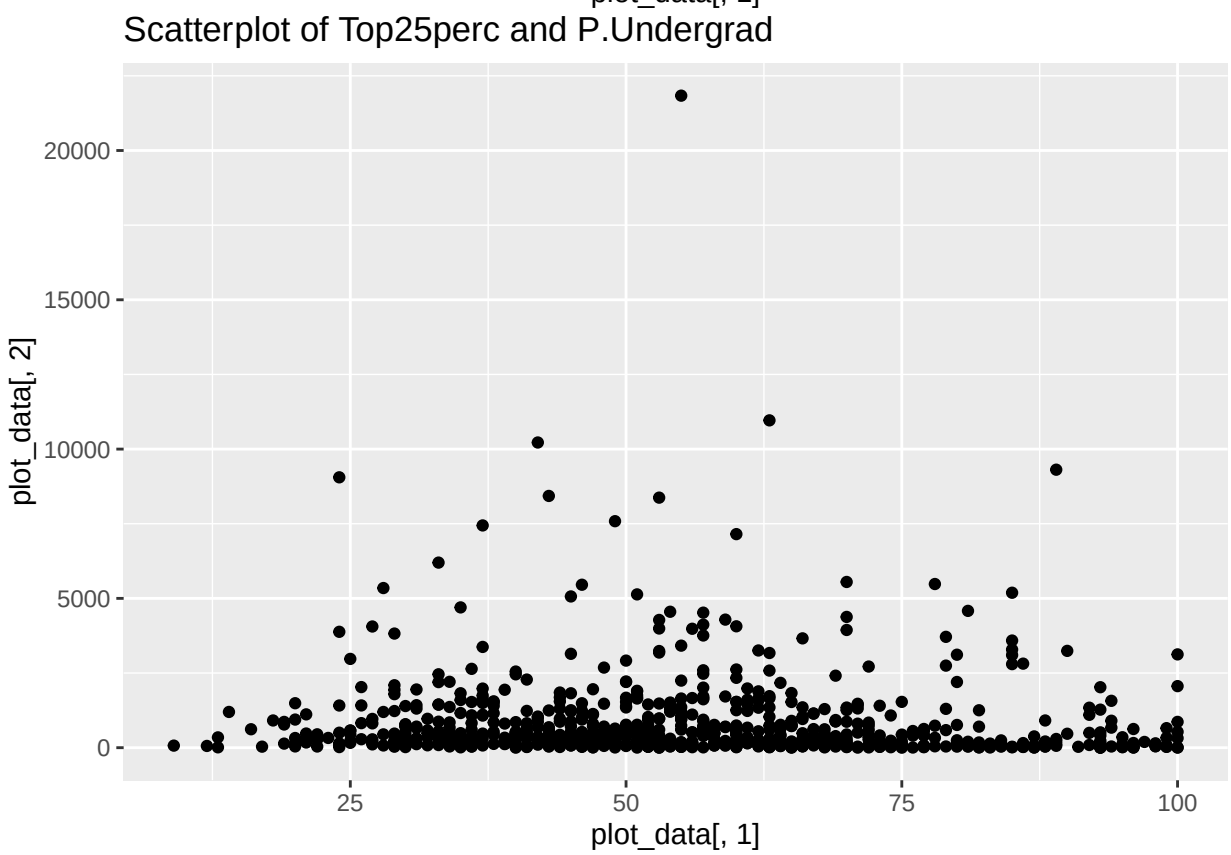

Scatterplot of Enroll and P.Undergrad
20000
15000
10000
5000
0
plot_data[, 2]
0
2000
4000
6000
plot_data[, 1]

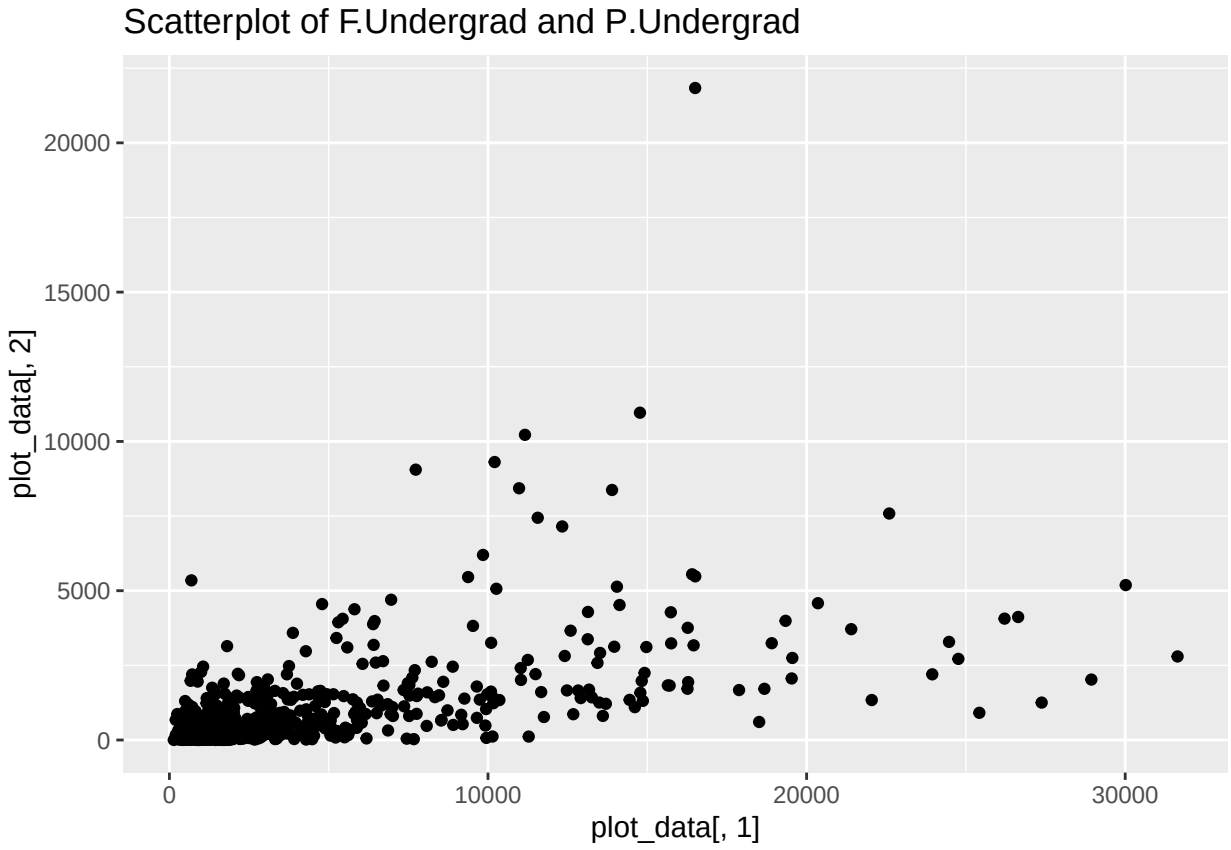

Scatterplot of Top10perc and Top25perc
100
75
50
25
plot_data[, 2]
0
25
50
75
100
plot_data[, 1]


Scatterplot of Top10perc and F.Undergrad
30000
20000
10000
0
plot_data[, 2]
0
25
50
75
100
plot_data[, 1]


Scatterplot of Top10perc and P.Undergrad
20000
15000
10000
5000
0
plot_data[, 2]
0
25
50
75
100
plot_data[, 1]


Scatterplot of Top25perc and F.Undergrad
30000
20000
10000
0
plot_data[, 2]
25
50
75
100
plot_data[, 1]


Scatterplot of Top25perc and P.Undergrad
20000
15000
10000
5000
0
plot_data[, 2]
25
50
75
100
plot_data[, 1]




```
sum(is.na(College))
```

Check for missing data

```
## [1] 0
```

Task 2: Linear regression using *all* variables

```
lm1 <- lm(Apps ~ ., data = College)
summary(lm1)
```

Model 1: Fit a linear regression model using all variables

```
##
## Call:
## lm(formula = Apps ~ ., data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4908.8  -430.2   -29.5   322.3  7852.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -445.08413  408.32855  -1.090 0.276053
## PrivateYes  -494.14897  137.81191  -3.586 0.000358 ***
## Accept         1.58581    0.04074  38.924  < 2e-16 ***
```

```
## Enroll        -0.88069    0.18596  -4.736 2.60e-06 ***
## Top10perc     49.92628    5.57824   8.950  < 2e-16 ***
## Top25perc    -14.23448    4.47914  -3.178 0.001543 **
## F.Undergrad    0.05739    0.03271   1.754 0.079785 .
## P.Undergrad    0.04445    0.03214   1.383 0.167114
## Outstate      -0.08587    0.01906  -4.506 7.64e-06 ***
## Room.Board     0.15103    0.04829   3.127 0.001832 **
## Books          0.02090    0.23841   0.088 0.930175
## Personal       0.03110    0.06308   0.493 0.622060
## PhD           -8.67850    4.63814  -1.871 0.061714 .
## Terminal      -3.33066    5.09494  -0.654 0.513492
## S.F.Ratio     15.38961   13.00622   1.183 0.237081
## perc.alumni    0.17867    4.10230   0.044 0.965273
## Expend         0.07790    0.01235   6.308 4.79e-10 ***
## Grad.Rate      8.66763    2.94893   2.939 0.003390 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1041 on 759 degrees of freedom
## Multiple R-squared:  0.9292, Adjusted R-squared:  0.9276
## F-statistic: 585.9 on 17 and 759 DF,  p-value: < 2.2e-16
```

Task 3: Linear regression using *meaningfull* variables

```
lm2 <- lm(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.Boarc
summary(lm2)
```

Model 2: Fit a linear regression model using selected variables

```
##
## Call:
## lm(formula = Apps ~ Private + Accept + Enroll + Top10perc + Top25perc +
##     F.Undergrad + Outstate + Room.Board + Books + Personal +
##     PhD + Terminal + S.F.Ratio + perc.alumni + Expend, data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4973.0  -400.4   -18.1   297.3  7791.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.638e+02  3.979e+02  -0.412 0.680623
## PrivateYes  -4.778e+02  1.380e+02  -3.461 0.000567 ***
## Accept       1.595e+00  4.064e-02  39.246  < 2e-16 ***
## Enroll      -8.845e-01  1.864e-01  -4.744 2.50e-06 ***
## Top10perc    5.041e+01  5.566e+00   9.057  < 2e-16 ***
## Top25perc   -1.323e+01  4.488e+00  -2.947 0.003306 **
## F.Undergrad  6.239e-02  3.174e-02   1.966 0.049688 *
## Outstate    -7.831e-02  1.895e-02  -4.132 4.00e-05 ***
## Room.Board   1.740e-01  4.791e-02   3.631 0.000301 ***
## Books        4.078e-03  2.395e-01   0.017 0.986419
## Personal     2.473e-02  6.271e-02   0.394 0.693454
## PhD         -7.802e+00  4.651e+00  -1.678 0.093842 .
## Terminal    -3.974e+00  5.115e+00  -0.777 0.437381
```

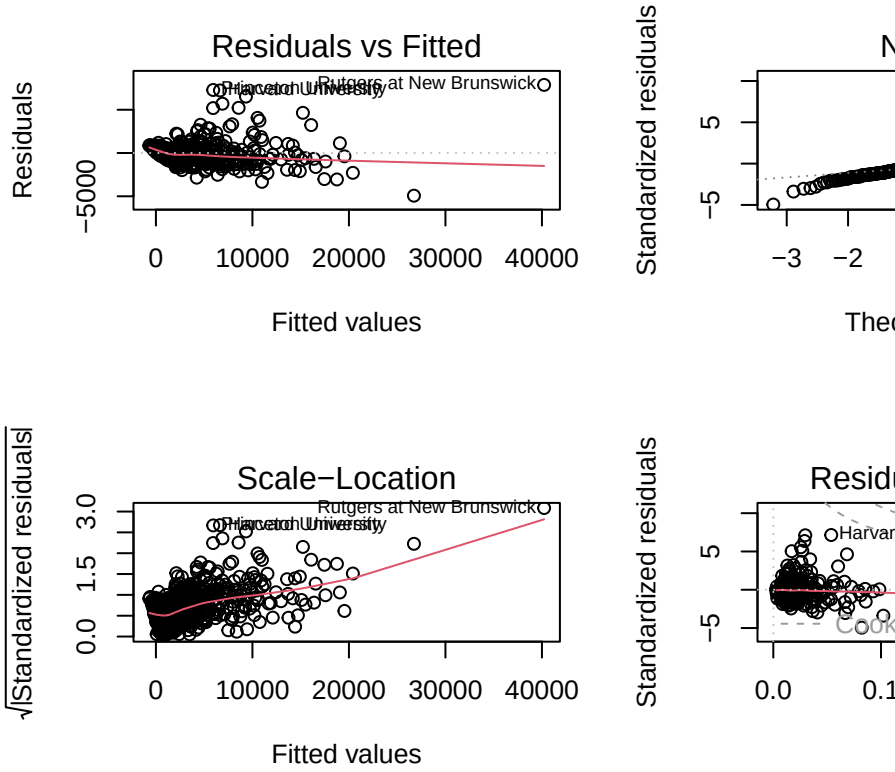
```
## S.F.Ratio    1.610e+01  1.307e+01   1.232 0.218423
## perc.alumni  2.417e+00  4.031e+00   0.600 0.548954
## Expend       7.525e-02  1.236e-02   6.091 1.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1047 on 761 degrees of freedom
## Multiple R-squared:  0.9283, Adjusted R-squared:  0.9269
## F-statistic: 656.7 on 15 and 761 DF,  p-value: < 2.2e-16
```

Task 4: Stepwise variable selection

```
stepwise.model <- step(lm(Apps ~ ., data = College), direction = "both", trace = FALSE)
```

Model 3: Perform stepwise selection using AIC as the selection criterion

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



Evaluate the final model's goodness of fit

Identify significant predictors

```
sig_preds <- names(which(summary(stepwise.model)$coefficients[,4] < 0.05))
```

```
coeffs <- coef(stepwise.model)
cat("Top10perc coefficient:", coeffs["Top10perc"], "\n")
```

Interpret the coefficients

```
## Top10perc coefficient: 50.41132
```

```r
cat("Private coefficient:", coeffs["Private"], "\n")
```

```
## Private coefficient: NA
```

Task 5: In-sample comparison of model 1,2 and 3

```r
library("caret")
```

```
## Loading required package: lattice
```

```r
train.control <- trainControl(method = "cv", number = 5)
mse1 <- train(Apps ~ ., data = College, method = "lm", trControl = train.control)
mse2 <- train(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.
mse3 <- train(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.
mse1$results$RMSE
```

```
## [1] 1097.76
```

```r
mse2$results$RMSE
```

```
## [1] 1120.517
```

```r
mse3$results$RMSE
```

```
## [1] 1167.075
```

Task 6: Out-of-sample comparison of model 1,2 and 3

```r
set.seed(123)
train.index <- createDataPartition(College$Apps, p = 0.8, list = FALSE)
train.data <- College[train.index, ]
test.data <- College[-train.index, ]
lm1.fit <- lm(Apps ~ ., data = train.data)
lm2.fit <- lm(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.
lm3.fit <- lm(Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.
lm1.pred <- predict(lm1.fit, newdata = test.data)
lm2.pred <- predict(lm2.fit, newdata = test.data)
lm3.pred <- predict(lm3.fit, newdata = test.data)
```

```r
train.mse <- c(mean((predict(lm1.fit, newdata = train.data) - train.data$Apps)^2),
               mean((predict(lm2.fit, newdata = train.data) - train.data$Apps)^2),
               mean((predict(lm3.fit, newdata = train.data) - train.data$Apps)^2))
cat("Training MSE:", train.mse, "\n")
```

Calculate training MSE

```
## Training MSE: 1037769 1053475 1083584
```

```r
test.mse <- c(mean((lm1.pred - test.data$Apps)^2),
              mean((lm2.pred - test.data$Apps)^2),
              mean((lm3.pred - test.data$Apps)^2))
cat("Test MSE:", test.mse, "\n")
```

Calculate test MSE

```
## Test MSE: 1224247 1231210 1371830
```

```
best.model.index <- which.min(test.mse)
```

Find index of the model with smallest test MSE

```
cat("Model", best.model.index, "is the best with a test MSE of", test.mse[best.model.index], "\n")
```

Output the best model

```
## Model 1 is the best with a test MSE of 1224247
```