> setwd("D:/Documents (Louis Booth)/R/data")

> pdf(file="boxplots.pdf")

> data.set = read.csv("College.csv", stringsAsFactors = FALSE)

> head(data.set)

X Private Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board

1 Abilene Christian University Yes 1660 1232 721 23 52 2885 537 7440 3300

2 Adelphi University Yes 2186 1924 512 16 29 2683 1227 12280 6450

3 Adrian College Yes 1428 1097 336 22 50 1036 99 11250 3750

4 Agnes Scott College Yes 417 349 137 60 89 510 63 12960 5450

5 Alaska Pacific University Yes 193 146 55 16 44 249 869 7560 4120

6 Albertson College Yes 587 479 158 38 62 678 41 13500 3335

Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate

1 450 2200 70 78 18.1 12 7041 60

2 750 1500 29 30 12.2 16 10527 56

3 400 1165 53 66 12.9 30 8735 54

4 450 875 92 97 7.7 37 19016 59

5 800 1500 76 72 11.9 2 10922 15

6 500 675 67 73 9.4 11 9727 55

> data\_1 <- data.set[data.set$Private == "Yes" | data.set$Private == "No",]

> data\_1$Private <- factor(data\_1$Private, levels=c("Yes", "No"))

> boxplot(data\_1$Personal~data\_1$Private)

> t.test(data\_1$Personal[data\_1$Private == "Yes"], data\_1$Personal[data\_1$Private == "No"], var.equal=FALSE)

Welch Two Sample t-test

data: data\_1$Personal[data\_1$Private == "Yes"] and data\_1$Personal[data\_1$Private == "No"]

t = -8.6277, df = 357.45, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-567.9730 -357.1078

sample estimates:

mean of x mean of y

1214.441 1676.981

> data\_2 <- read.csv("College.csv", stringsAsFactors = FALSE)

> rank <- order(data\_2$Top10perc, decreasing=TRUE)

> data\_2 <- data\_2[rank,]

> data\_2$Group[1:259] <- "Top"

> data\_2$Group[260:518] <- "Middle"

> data\_2$Group[519:777] <- "Bottom"

> data\_2$Group <- factor(data\_2$Group, levels=c("Top", "Middle", "Bottom"))

> str(data\_2)

'data.frame': 777 obs. of 20 variables:

$ X : chr "Massachusetts Institute of Technology" "Harvey Mudd College" "University of California at Berkeley" "Yale University" ...

$ Private : chr "Yes" "Yes" "No" "Yes" ...

$ Apps : int 6411 1377 19873 10705 13789 13865 13218 7837 12586 8587 ...

$ Accept : int 2140 572 8252 2453 3893 2165 2042 4527 3239 2273 ...

$ Enroll : int 1078 178 3215 1317 1583 1606 1153 2276 1462 1087 ...

$ Top10perc : int 96 95 95 95 90 90 90 89 87 87 ...

$ Top25perc : int 99 100 100 99 98 100 98 99 95 99 ...

$ F.Undergrad: int 4481 654 19532 5217 6188 6862 4540 8528 5643 3918 ...

$ P.Undergrad: int 28 5 2061 83 53 320 146 654 349 32 ...

$ Outstate : int 20100 17230 11648 19840 18590 18485 19900 6489 19528 19545 ...

$ Room.Board : int 5975 6690 6246 6510 5950 6410 5910 4438 5926 6070 ...

$ Books : int 725 700 636 630 625 500 675 795 720 550 ...

$ Personal : int 1600 900 1933 2115 1162 1920 1575 1164 1100 1100 ...

$ PhD : int 99 100 93 96 95 97 91 92 99 95 ...

$ Terminal : int 99 100 97 96 96 97 96 92 100 99 ...

$ S.F.Ratio : num 10.1 8.2 15.8 5.8 5 9.9 8.4 19.3 7.6 4.7 ...

$ perc.alumni: int 35 46 10 49 44 52 54 33 39 49 ...

$ Expend : int 33541 21569 13919 40386 27206 37219 28320 11271 20440 29619 ...

$ Grad.Rate : int 94 100 78 99 97 100 99 70 97 98 ...

$ Group : Factor w/ 3 levels "Top","Middle",..: 1 1 1 1 1 1 1 1 1 1 ...

> boxplot(data\_2$Grad.Rate~data\_2$Group)

> anova(lm(data\_2$Grad.Rate~data\_2$Group))

Analysis of Variance Table

Response: data\_2$Grad.Rate

Df Sum Sq Mean Sq F value Pr(>F)

data\_2$Group 2 44574 22286.9 93.545 < 2.2e-16 \*\*\*

Residuals 774 184403 238.2

---

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

> pairwise.t.test(data\_2$Grad.Rate, data\_2$Group, p.adj="bonferroni")

Pairwise comparisons using t tests with pooled SD

data: data\_2$Grad.Rate and data\_2$Group

Top Middle

Middle 5.4e-14 -

Bottom < 2e-16 2.7e-08

P value adjustment method: bonferroni

> dev.off()

null device

1

Based on the p-values reported above, it is clear that all three means are different from one another. This means “Top” mean is different from “Middle” mean, “Top” mean is different from “Bottom” mean, and “Bottom” mean is different from “Middle” mean.