Optional week 3 assignment

# Question: Based on your analysis, would you conclude that there is a significant association between college major category and income?

create data frame with major\_category and median earnings per row

cat <- college$major\_category  
earn <- college$median  
df <- data.frame(cat, earn)  
head(df)

## cat earn  
## 1 Engineering 40000  
## 2 Engineering 37000  
## 3 Engineering 45000  
## 4 Engineering 35000  
## 5 Engineering 62000  
## 6 Engineering 44700

provide aggregate means per category and sort to see signifcant changes between categories

agg <- aggregate(df$earn ~ df$cat, df, mean)  
agg <- agg[order(agg$`df$earn`, decreasing = TRUE),]  
agg

## df$cat df$earn  
## 4 Business 49153.85  
## 3 Biology & Life Science 43864.29  
## 1 Agriculture & Natural Resources 43500.00  
## 5 Communications & Journalism 42000.00  
## 11 Industrial Arts & Consumer Services 40428.57  
## 14 Physical Sciences 40400.00  
## 8 Engineering 40393.10  
## 9 Health 40316.67  
## 15 Psychology & Social Work 39888.89  
## 16 Social Science 39066.67  
## 2 Arts 38050.00  
## 7 Education 37937.50  
## 13 Law & Public Policy 37800.00  
## 10 Humanities & Liberal Arts 35166.67  
## 6 Computers & Mathematics 34718.18  
## 12 Interdisciplinary 27500.00

Relevel the factors for the lowest income (“Interdisciplinary”) and fit a model for the data

df$cat <- relevel(df$cat, ref = "Interdisciplinary")  
fit <- lm(df$earn ~ df$cat)  
summary(fit)

##   
## Call:  
## lm(formula = df$earn ~ df$cat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20154 -6800 -2893 4833 60846   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 27500 11355 2.422  
## df$catAgriculture & Natural Resources 16000 11909 1.343  
## df$catArts 10550 12044 0.876  
## df$catBiology & Life Science 16364 11754 1.392  
## df$catBusiness 21654 11784 1.838  
## df$catCommunications & Journalism 14500 12696 1.142  
## df$catComputers & Mathematics 7218 11860 0.609  
## df$catEducation 10438 11705 0.892  
## df$catEngineering 12893 11549 1.116  
## df$catHealth 12817 11819 1.084  
## df$catHumanities & Liberal Arts 7667 11728 0.654  
## df$catIndustrial Arts & Consumer Services 12929 12139 1.065  
## df$catLaw & Public Policy 10300 12439 0.828  
## df$catPhysical Sciences 12900 11909 1.083  
## df$catPsychology & Social Work 12389 11969 1.035  
## df$catSocial Science 11567 11969 0.966  
## Pr(>|t|)   
## (Intercept) 0.0166 \*  
## df$catAgriculture & Natural Resources 0.1811   
## df$catArts 0.3824   
## df$catBiology & Life Science 0.1658   
## df$catBusiness 0.0680 .  
## df$catCommunications & Journalism 0.2551   
## df$catComputers & Mathematics 0.5437   
## df$catEducation 0.3739   
## df$catEngineering 0.2660   
## df$catHealth 0.2798   
## df$catHumanities & Liberal Arts 0.5142   
## df$catIndustrial Arts & Consumer Services 0.2885   
## df$catLaw & Public Policy 0.4089   
## df$catPhysical Sciences 0.2804   
## df$catPsychology & Social Work 0.3022   
## df$catSocial Science 0.3354   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11360 on 157 degrees of freedom  
## Multiple R-squared: 0.1054, Adjusted R-squared: 0.01995   
## F-statistic: 1.233 on 15 and 157 DF, p-value: 0.2522

Summary shows that there is no signifcant change in income for categories as compared to the lowest income level

Boxplot of categories

boxplot(df$earn ~ df$cat, las = 2)

