ANLY 500 Final Project

# Read Data

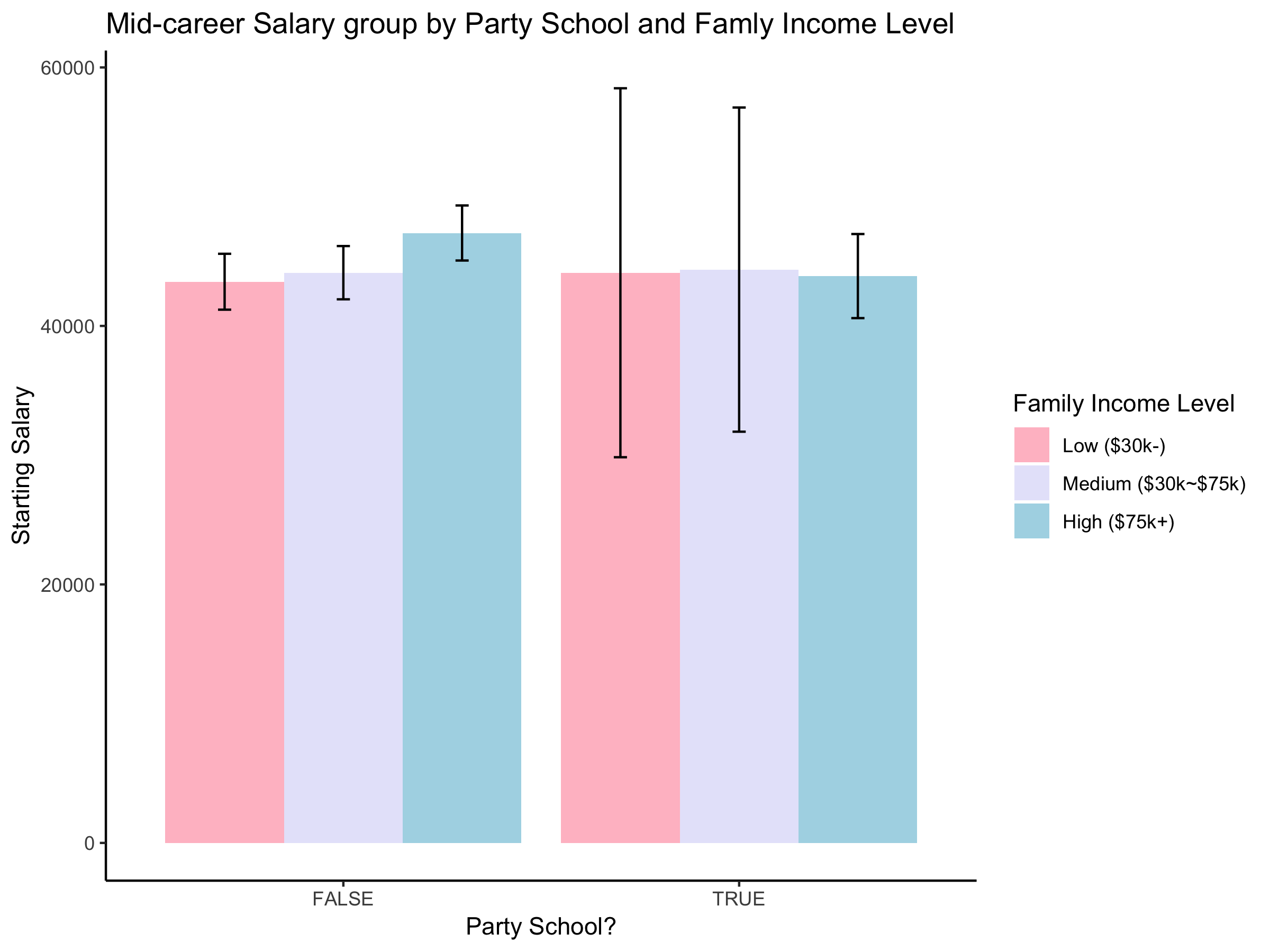
data = read\_csv('data\_cleaned.csv', col\_types = paste0('cfflfff', strrep('i', 22), strrep('d', 8)))  
data$Tier = factor(data$Tier, levels = c("Selective private", "Selective public", "Highly selective private", "Highly selective public", "Other elite schools (public and private)", "Ivy Plus"))

# Algorithm and Models

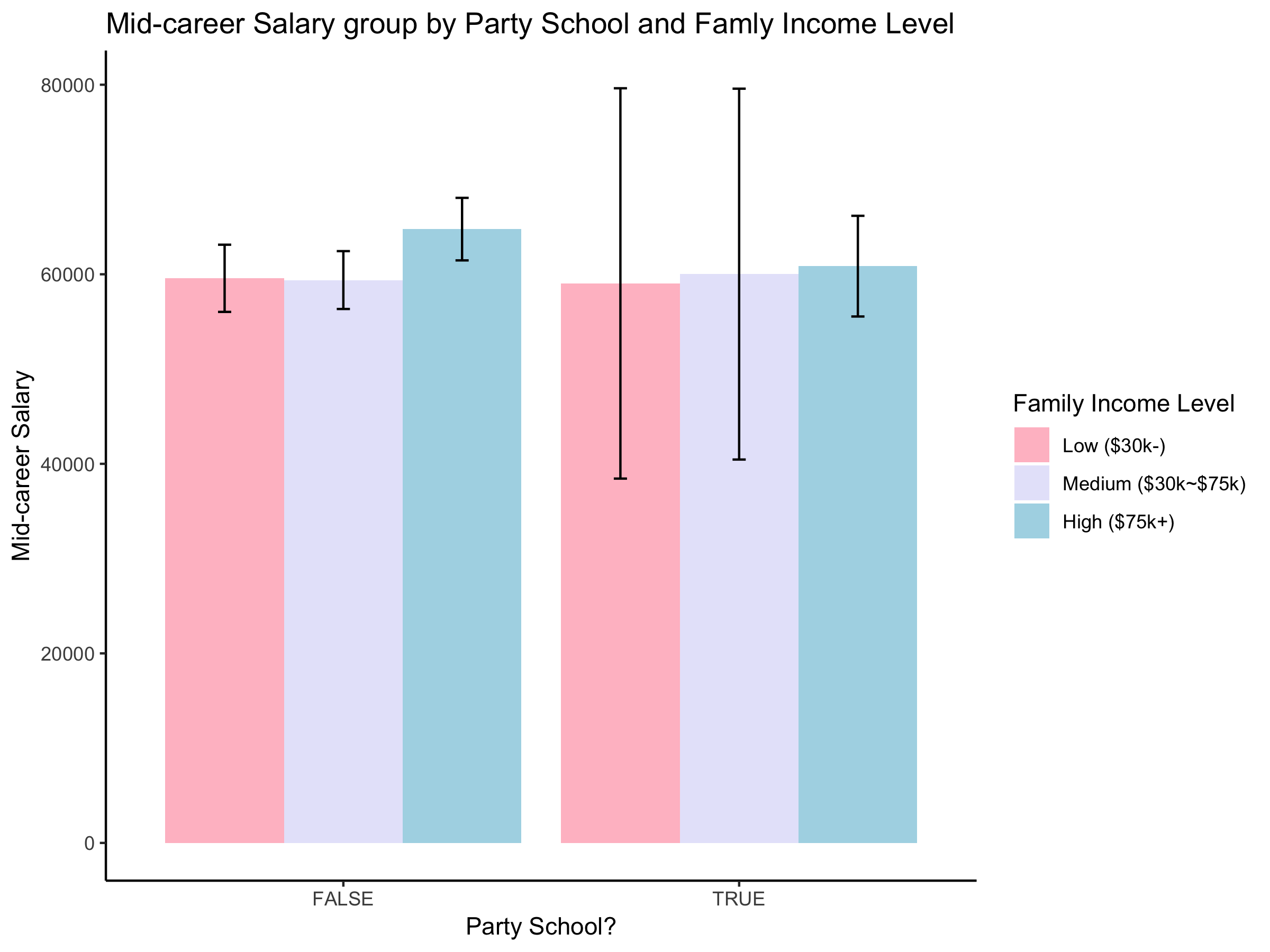
# Gender & Family

data\_gender\_p6 = data %>%  
 melt(id.vars=1:8, measure.vars=c( "MN\_EARN\_WNE\_MALE0\_P6", "MN\_EARN\_WNE\_MALE1\_P6"), variable.name="Gender", value.name="Income") %>%  
 mutate(  
 Gender = factor(Gender, levels = c("MN\_EARN\_WNE\_MALE0\_P6", "MN\_EARN\_WNE\_MALE1\_P6"), labels = c("Female", "Male")),  
 ROI = (Income-COSTT4\_A)/COSTT4\_A\*100  
 )  
   
data\_gender\_p10 = data %>%  
 melt(id.vars=1:8, measure.vars=c( "MN\_EARN\_WNE\_MALE0\_P10", "MN\_EARN\_WNE\_MALE1\_P10"), variable.name="Gender", value.name="Income") %>%  
 mutate(  
 Gender = factor(Gender, levels = c("MN\_EARN\_WNE\_MALE0\_P10", "MN\_EARN\_WNE\_MALE1\_P10"), labels = c("Female", "Male")),  
 ROI = (Income-COSTT4\_A)/COSTT4\_A\*100  
 )  
  
data\_family\_p6 = data %>%  
 melt(id.vars=1:8, measure.vars=c( "MN\_EARN\_WNE\_INC1\_P6", "MN\_EARN\_WNE\_INC2\_P6", "MN\_EARN\_WNE\_INC3\_P6"), variable.name="Family.Income", value.name="Income") %>%  
 mutate(  
 Family.Income = factor(Family.Income, levels = c("MN\_EARN\_WNE\_INC1\_P6", "MN\_EARN\_WNE\_INC2\_P6", "MN\_EARN\_WNE\_INC3\_P6"), labels = c("Low", "Middle", "High")),  
 ROI = (Income-COSTT4\_A)/COSTT4\_A\*100  
 )  
  
data\_family\_p10 = data %>%  
 melt(id.vars=1:8, measure.vars=c( "MN\_EARN\_WNE\_INC1\_P10", "MN\_EARN\_WNE\_INC2\_P10", "MN\_EARN\_WNE\_INC3\_P10"), variable.name="Family.Income", value.name="Income") %>%  
 mutate(  
 Family.Income = factor(Family.Income, levels = c("MN\_EARN\_WNE\_INC1\_P10", "MN\_EARN\_WNE\_INC2\_P10", "MN\_EARN\_WNE\_INC3\_P10"), labels = c("Low", "Middle", "High")),  
 ROI = (Income-COSTT4\_A)/COSTT4\_A\*100  
 )

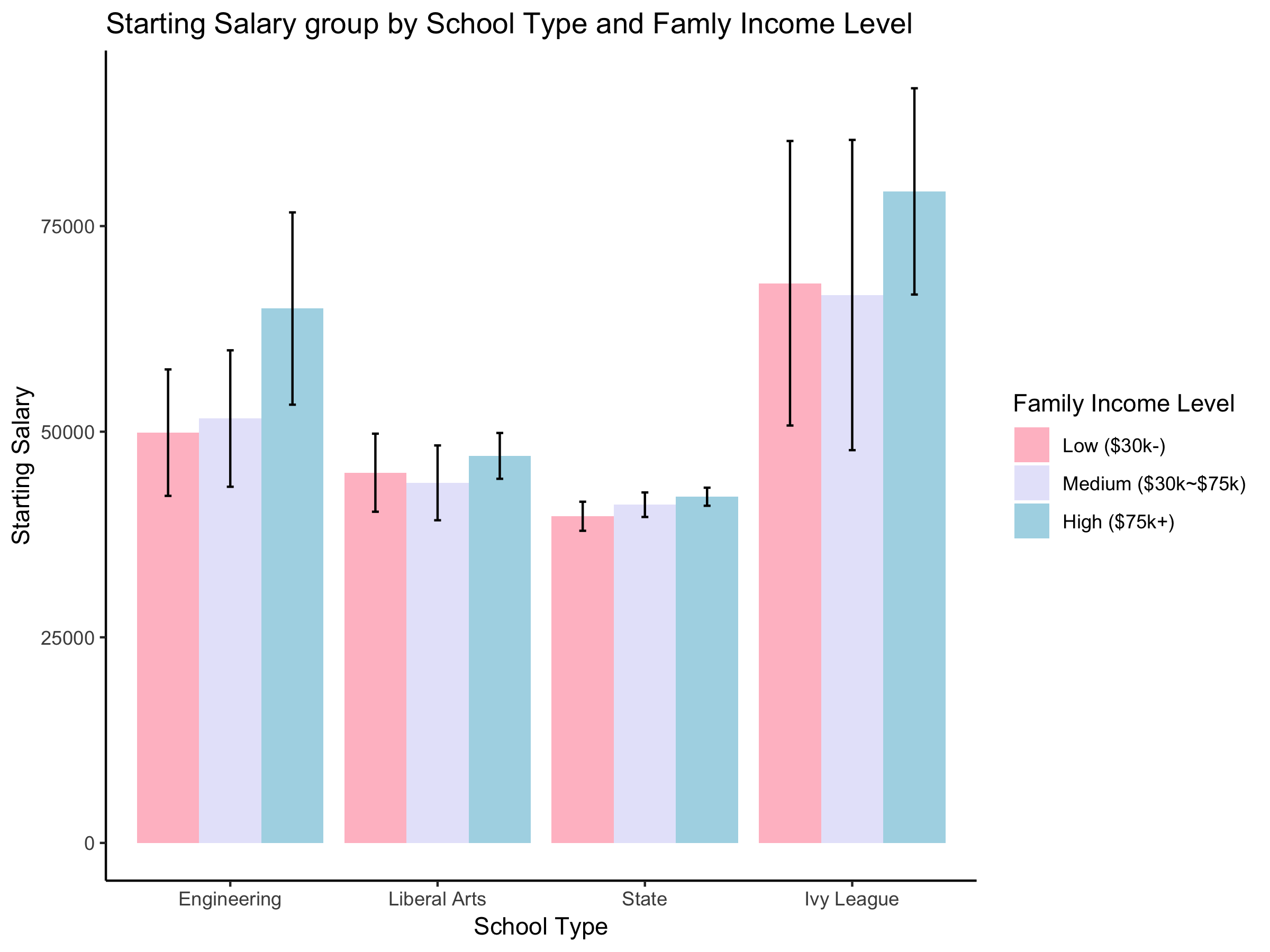
#Explore the interaction effect with what you can't change - family and gender  
plot\_elements = function() {  
 list(  
 theme\_classic(),  
 stat\_summary(fun.y = mean,  
 geom = "bar",  
 position = "dodge"),  
 stat\_summary(fun.data = mean\_cl\_normal,  
 geom = "errorbar",  
 position = position\_dodge(width = 0.9),  
 width = 0.1)  
 )  
}  
  
data\_family\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(Is.Party, Income, fill=Family.Income)) +  
 xlab("Party School?")+  
 ylab("Starting Salary") +  
 ggtitle("Mid-career Salary group by Party School and Famly Income Level") +  
 scale\_fill\_manual(name = "Family Income Level",  
 labels = c("Low ($30k-)", "Medium ($30k~$75k)", "High ($75k+)"),  
 values = c("Pink", "Lavender", "LightBlue")) +  
 plot\_elements()



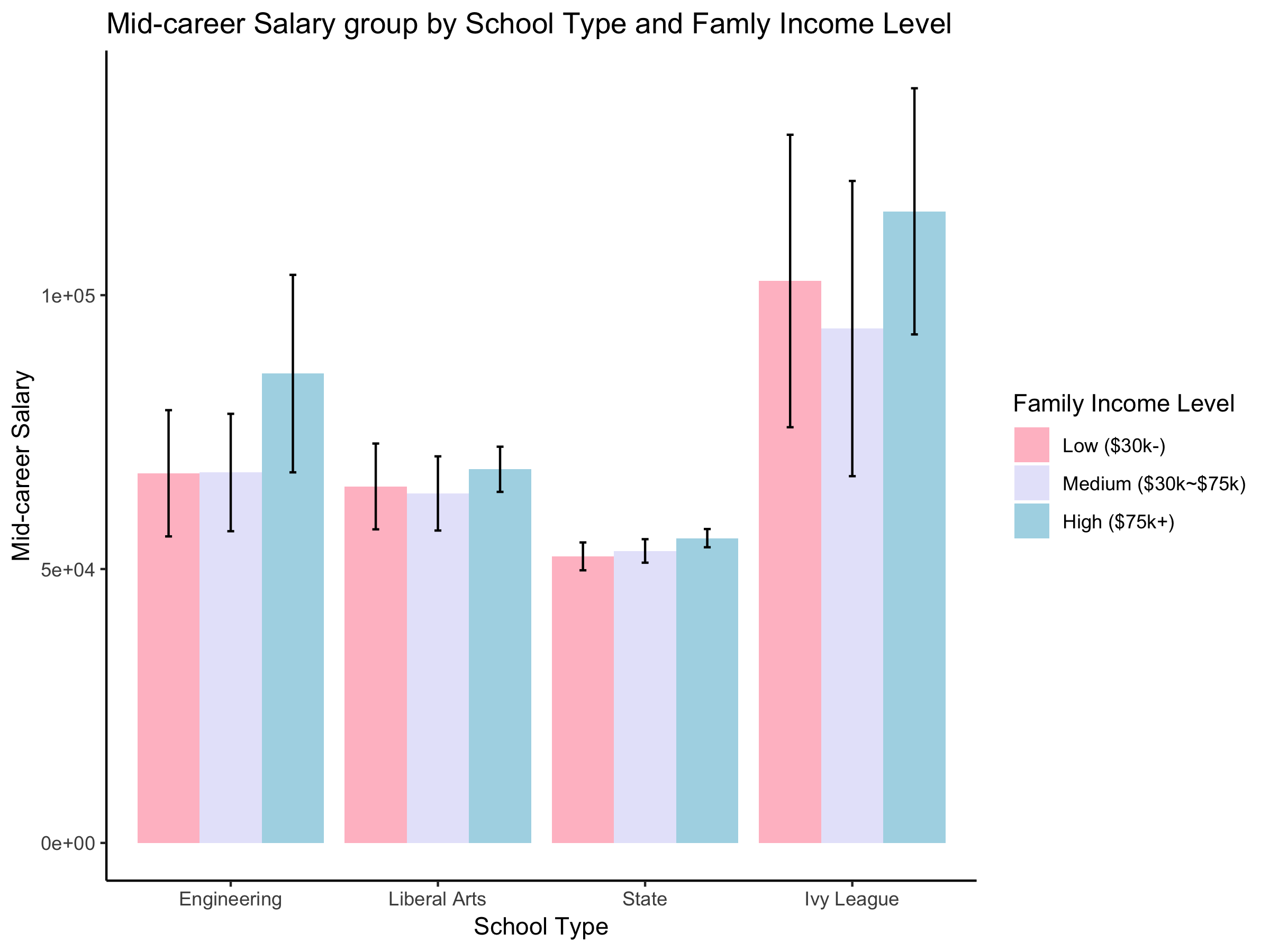
data\_family\_p10 %>%  
 drop\_na() %>%  
 ggplot(aes(Is.Party, Income, fill=Family.Income)) +  
 xlab("Party School?")+  
 ylab("Mid-career Salary") +  
 ggtitle("Mid-career Salary group by Party School and Famly Income Level") +  
 scale\_fill\_manual(name = "Family Income Level",  
 labels = c("Low ($30k-)", "Medium ($30k~$75k)", "High ($75k+)"),  
 values = c("Pink", "Lavender", "LightBlue")) +  
 plot\_elements()



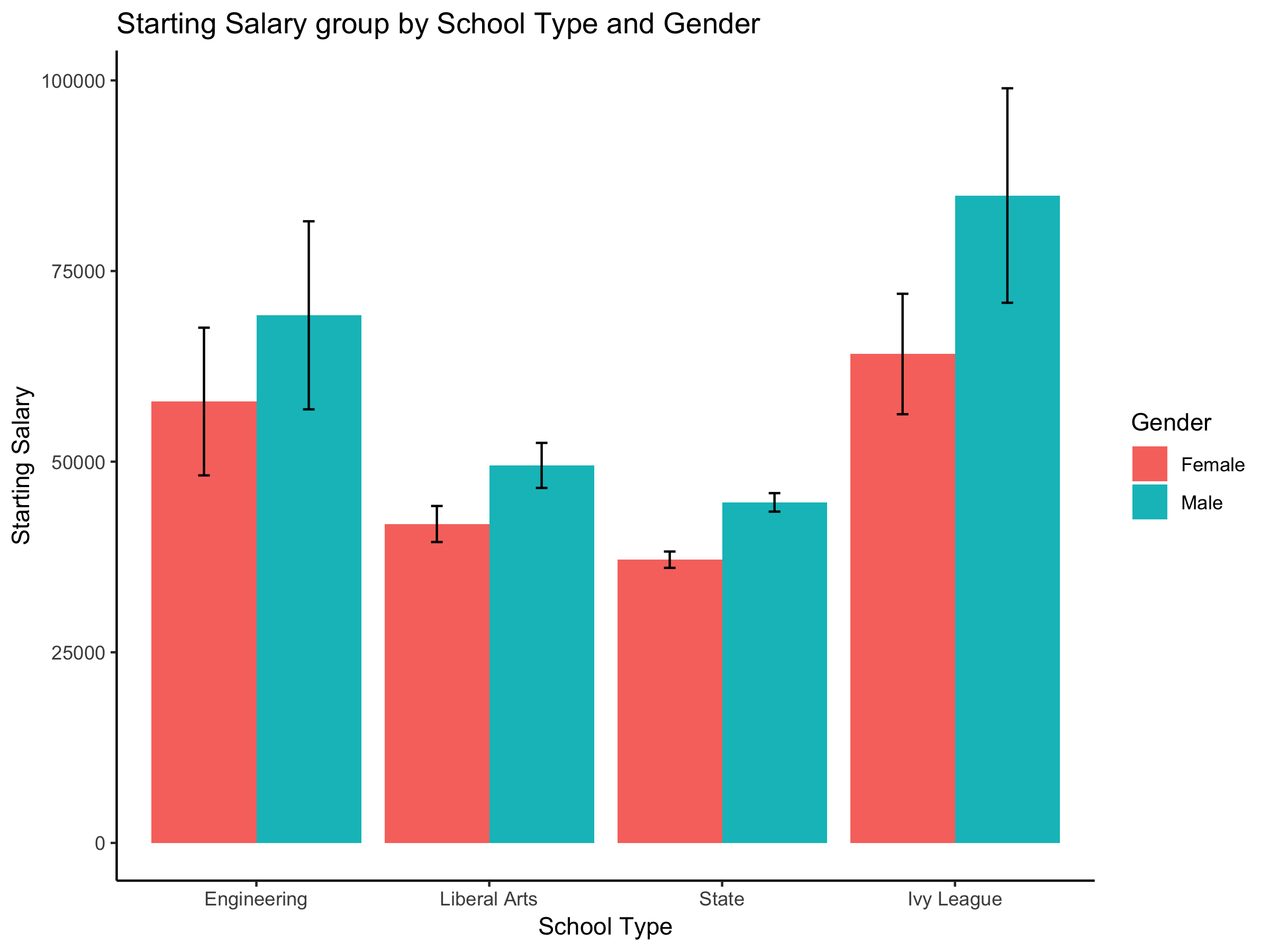
data\_family\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(School.Type, Income, fill=Family.Income)) +  
 xlab("School Type")+  
 ylab("Starting Salary") +  
 ggtitle("Starting Salary group by School Type and Famly Income Level") +  
 scale\_fill\_manual(name = "Family Income Level",  
 labels = c("Low ($30k-)", "Medium ($30k~$75k)", "High ($75k+)"),  
 values = c("Pink", "Lavender", "LightBlue")) +  
 plot\_elements()



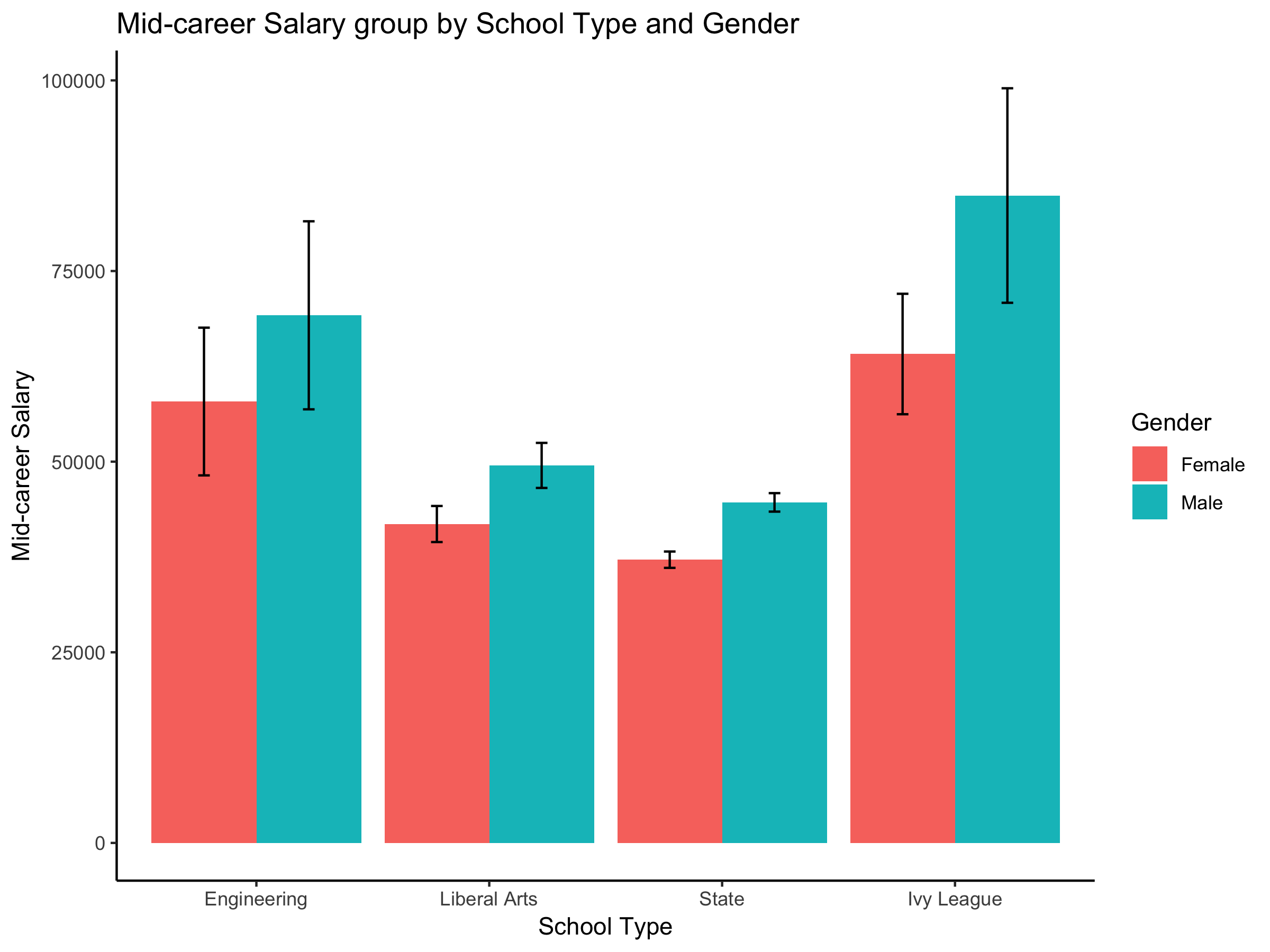
data\_family\_p10 %>%  
 drop\_na() %>%  
 ggplot(aes(School.Type, Income, fill=Family.Income)) +  
 xlab("School Type")+  
 ylab("Mid-career Salary") +  
 ggtitle("Mid-career Salary group by School Type and Famly Income Level") +  
 scale\_fill\_manual(name = "Family Income Level",  
 labels = c("Low ($30k-)", "Medium ($30k~$75k)", "High ($75k+)"),  
 values = c("Pink", "Lavender", "LightBlue")) +  
 plot\_elements()



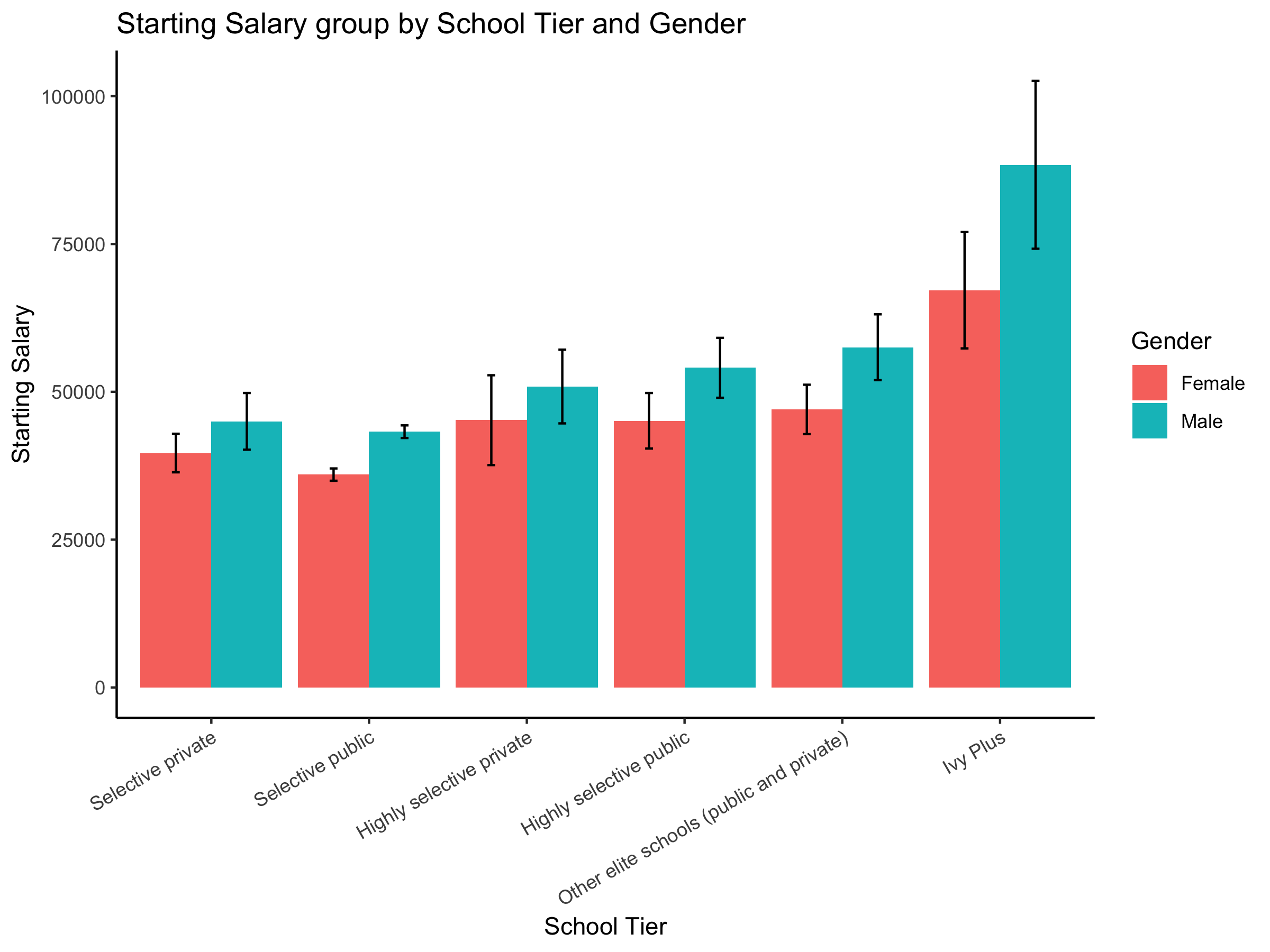
data\_gender\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(School.Type, Income, fill=Gender)) +  
 xlab("School Type")+  
 ylab("Starting Salary") +  
 ggtitle("Starting Salary group by School Type and Gender") +  
 plot\_elements()



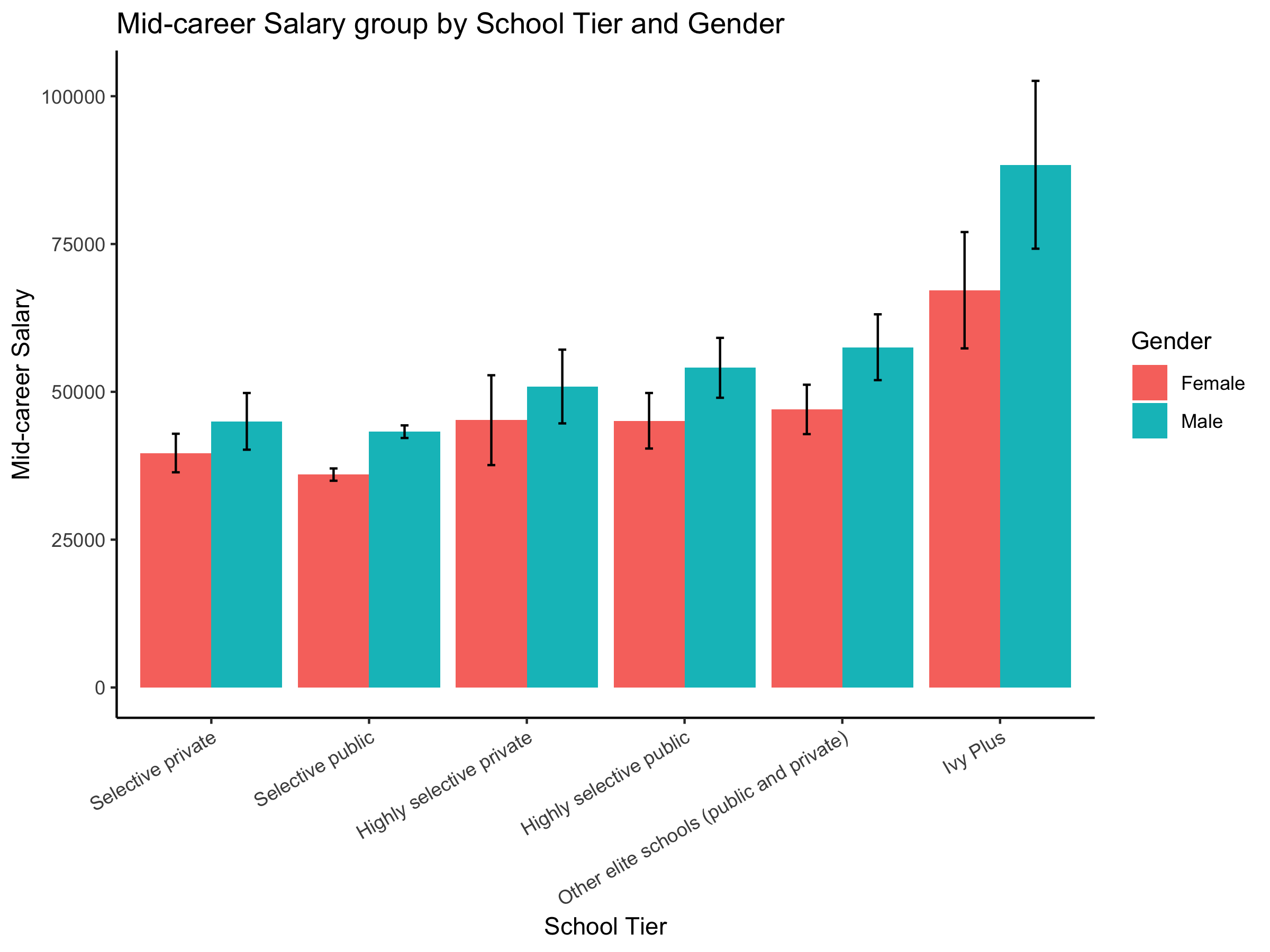
data\_gender\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(School.Type, Income, fill=Gender)) +  
 xlab("School Type")+  
 ylab("Mid-career Salary") +  
 ggtitle("Mid-career Salary group by School Type and Gender") +  
 plot\_elements()



data\_gender\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(Tier, Income, fill=Gender)) +  
 xlab("School Tier")+  
 ylab("Starting Salary") +  
 ggtitle("Starting Salary group by School Tier and Gender") +  
 plot\_elements() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



data\_gender\_p6 %>%  
 drop\_na() %>%  
 ggplot(aes(Tier, Income, fill=Gender)) +  
 xlab("School Tier")+  
 ylab("Mid-career Salary") +  
 ggtitle("Mid-career Salary group by School Tier and Gender") +  
 plot\_elements() +   
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



#T-test to see if there is a difference in starting income from the two different gender group  
#H0: Male = Female  
#H1: Male <> Female  
data\_gender\_p6 = drop\_na(data\_gender\_p6)  
t.test(data\_gender\_p6$Income ~ data\_gender\_p6$Gender)

##   
## Welch Two Sample t-test  
##   
## data: data\_gender\_p6$Income by data\_gender\_p6$Gender  
## t = -5.8939, df = 268.79, p-value = 1.127e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -11286.586 -5634.247  
## sample estimates:  
## mean in group Female mean in group Male   
## 41504.17 49964.58

M = tapply(data\_gender\_p6$Income, data\_gender\_p6$Gender, mean)   
N = tapply(data\_gender\_p6$Income, data\_gender\_p6$Gender, length)   
stdev = tapply(data\_gender\_p6$Income, data\_gender\_p6$Gender, sd)   
M;stdev;N

## Female Male   
## 41504.17 49964.58

## Female Male   
## 10527.05 13634.46

## Female Male   
## 144 144

effect1 = d.ind.t(  
 m1 = M[1], m2 = M[2],  
 sd1 = stdev[1], sd2 = stdev[2],   
 n1 = N[1], n2 = N[2], a = .05)   
effect1$d

## Female   
## -0.6946008

pwr.t.test(n = NULL, d = effect1$d,   
 sig.level = .05,  
 power = .80, type = "two.sample",   
 alternative = "two.sided")

##   
## Two-sample t test power calculation   
##   
## n = 33.52412  
## d = 0.6946008  
## sig.level = 0.05  
## power = 0.8  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

#T-test to see if there is a difference in mid-career income from the two different gender group  
#H0: Male = Female  
#H1: Male <> Female  
data\_gender\_p10 = drop\_na(data\_gender\_p10)  
t.test(data\_gender\_p10$Income ~ data\_gender\_p10$Gender)

##   
## Welch Two Sample t-test  
##   
## data: data\_gender\_p10$Income by data\_gender\_p10$Gender  
## t = -7.2668, df = 254.21, p-value = 4.515e-12  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -20421.40 -11712.87  
## sample estimates:  
## mean in group Female mean in group Male   
## 54377.62 70444.76

M = tapply(data\_gender\_p10$Income, data\_gender\_p10$Gender, mean)   
N = tapply(data\_gender\_p10$Income, data\_gender\_p10$Gender, length)   
stdev = tapply(data\_gender\_p10$Income, data\_gender\_p10$Gender, sd)   
M;stdev;N

## Female Male   
## 54377.62 70444.76

## Female Male   
## 15161.61 21661.07

## Female Male   
## 143 143

effect2 = d.ind.t(  
 m1 = M[1], m2 = M[2],  
 sd1 = stdev[1], sd2 = stdev[2],   
 n1 = N[1], n2 = N[2], a = .05)   
effect2$d

## Female   
## -0.8593915

pwr.t.test(n = NULL, d = effect2$d,   
 sig.level = .05,  
 power = .80, type = "two.sample",   
 alternative = "two.sided")

##   
## Two-sample t test power calculation   
##   
## n = 22.2571  
## d = 0.8593915  
## sig.level = 0.05  
## power = 0.8  
## alternative = two.sided  
##   
## NOTE: n is number in \*each\* group

model\_gender\_p6\_1 = lm(ROI ~ Gender, data = data\_gender\_p6)  
summary(model\_gender\_p6\_1)

##   
## Call:  
## lm(formula = ROI ~ Gender, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -113.21 -56.15 10.65 47.45 163.43   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.011 4.910 7.946 4.48e-14 \*\*\*  
## GenderMale 28.513 6.944 4.106 5.25e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 58.92 on 286 degrees of freedom  
## Multiple R-squared: 0.05568, Adjusted R-squared: 0.05238   
## F-statistic: 16.86 on 1 and 286 DF, p-value: 5.25e-05

model\_gender\_p6\_2 = lm(ROI ~ Gender + CONTROL, data = data\_gender\_p6)  
summary(model\_gender\_p6\_2)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.724 -19.943 -3.285 16.451 124.007   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -22.943 3.548 -6.466 4.35e-10 \*\*\*  
## GenderMale 28.513 3.755 7.593 4.48e-13 \*\*\*  
## CONTROLPublic 101.380 3.851 26.324 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 31.86 on 285 degrees of freedom  
## Multiple R-squared: 0.7248, Adjusted R-squared: 0.7229   
## F-statistic: 375.3 on 2 and 285 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_1, model\_gender\_p6\_2)

## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender  
## Model 2: ROI ~ Gender + CONTROL  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 286 992796   
## 2 285 289330 1 703466 692.94 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model\_gender\_p6\_3 = lm(ROI ~ Gender + CONTROL + School.Type, data = data\_gender\_p6)  
summary(model\_gender\_p6\_3)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL + School.Type, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.540 -18.522 -0.978 14.844 124.190   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.576 6.880 1.392 0.165   
## GenderMale 28.513 3.471 8.215 7.73e-15 \*\*\*  
## CONTROLPublic 80.737 10.760 7.503 8.18e-13 \*\*\*  
## School.TypeLiberal Arts -44.163 7.281 -6.065 4.22e-09 \*\*\*  
## School.TypeState -12.060 9.699 -1.243 0.215   
## School.TypeIvy League -1.736 10.798 -0.161 0.872   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.45 on 282 degrees of freedom  
## Multiple R-squared: 0.7674, Adjusted R-squared: 0.7632   
## F-statistic: 186 on 5 and 282 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_2, model\_gender\_p6\_3)

## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender + CONTROL  
## Model 2: ROI ~ Gender + CONTROL + School.Type  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 285 289330   
## 2 282 244589 3 44742 17.195 2.78e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model\_gender\_p6\_4 = lm(ROI ~ Gender + CONTROL + School.Type + STABBR, data = data\_gender\_p6)  
summary(model\_gender\_p6\_4)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL + School.Type + STABBR, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.723 -13.174 0.123 12.095 124.008   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.585 8.928 1.298 0.19567   
## GenderMale 28.513 3.223 8.846 < 2e-16 \*\*\*  
## CONTROLPublic 79.272 14.630 5.418 1.45e-07 \*\*\*  
## School.TypeLiberal Arts -43.020 7.760 -5.544 7.70e-08 \*\*\*  
## School.TypeState -12.422 13.451 -0.923 0.35668   
## School.TypeIvy League 7.513 11.903 0.631 0.52851   
## STABBRMA 7.798 10.288 0.758 0.44920   
## STABBRAL -9.958 14.302 -0.696 0.48693   
## STABBRAR 7.799 19.787 0.394 0.69381   
## STABBRCO -1.816 15.339 -0.118 0.90588   
## STABBRCT -13.454 12.608 -1.067 0.28697   
## STABBRDE 14.312 19.787 0.723 0.47018   
## STABBRFL 18.549 9.610 1.930 0.05473 .   
## STABBRGA -8.245 14.302 -0.577 0.56479   
## STABBRID -7.146 11.925 -0.599 0.54960   
## STABBRIL -31.088 12.110 -2.567 0.01085 \*   
## STABBRIN -15.191 14.435 -1.052 0.29368   
## STABBRIA 19.617 11.963 1.640 0.10235   
## STABBRKS 17.494 11.925 1.467 0.14368   
## STABBRKY -15.338 14.302 -1.072 0.28457   
## STABBRLA 45.950 19.787 2.322 0.02105 \*   
## STABBRME 2.598 12.833 0.202 0.83977   
## STABBRMI -14.212 10.538 -1.349 0.17869   
## STABBRMN -3.048 12.833 -0.237 0.81249   
## STABBRMS -13.973 14.302 -0.977 0.32955   
## STABBRMT 29.418 19.787 1.487 0.13838   
## STABBRNV 19.954 14.302 1.395 0.16424   
## STABBRNH -17.044 22.541 -0.756 0.45031   
## STABBRNJ -10.001 16.037 -0.624 0.53345   
## STABBRNM 37.398 23.837 1.569 0.11797   
## STABBRNY -1.845 8.000 -0.231 0.81781   
## STABBRNC 1.892 8.917 0.212 0.83215   
## STABBRND 29.410 19.787 1.486 0.13849   
## STABBROH -16.855 9.936 -1.696 0.09110 .   
## STABBROR -14.609 9.707 -1.505 0.13363   
## STABBRPA 3.178 8.669 0.367 0.71426   
## STABBRRI -28.379 15.240 -1.862 0.06380 .   
## STABBRSC -12.182 19.787 -0.616 0.53871   
## STABBRTN -31.492 12.684 -2.483 0.01371 \*   
## STABBRTX 14.031 14.302 0.981 0.32752   
## STABBRUT 26.155 11.925 2.193 0.02924 \*   
## STABBRVT -25.836 14.435 -1.790 0.07473 .   
## STABBRVA -6.139 9.509 -0.646 0.51919   
## STABBRWA -16.241 9.707 -1.673 0.09560 .   
## STABBRWY 53.381 19.787 2.698 0.00747 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.35 on 243 degrees of freedom  
## Multiple R-squared: 0.8271, Adjusted R-squared: 0.7958   
## F-statistic: 26.42 on 44 and 243 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_3, model\_gender\_p6\_4)

## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender + CONTROL + School.Type  
## Model 2: ROI ~ Gender + CONTROL + School.Type + STABBR  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 282 244589   
## 2 243 181757 39 62832 2.1539 0.0002424 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Is.Party is not helpful  
model\_gender\_p6\_5 = lm(ROI ~ Gender + CONTROL + School.Type + STABBR + Is.Party, data = data\_gender\_p6)  
summary(model\_gender\_p6\_5)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL + School.Type + STABBR +   
## Is.Party, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.813 -13.161 0.121 12.358 123.918   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.650 8.951 1.302 0.1943   
## GenderMale 28.513 3.229 8.829 < 2e-16 \*\*\*  
## CONTROLPublic 79.141 14.672 5.394 1.64e-07 \*\*\*  
## School.TypeLiberal Arts -43.001 7.776 -5.530 8.28e-08 \*\*\*  
## School.TypeState -12.266 13.497 -0.909 0.3643   
## School.TypeIvy League 7.529 11.926 0.631 0.5284   
## STABBRMA 7.723 10.314 0.749 0.4547   
## STABBRAL -10.048 14.336 -0.701 0.4841   
## STABBRAR 7.709 19.831 0.389 0.6978   
## STABBRCO -1.806 15.369 -0.118 0.9065   
## STABBRCT -13.540 12.639 -1.071 0.2851   
## STABBRDE 14.222 19.831 0.717 0.4739   
## STABBRFL 19.238 10.141 1.897 0.0590 .   
## STABBRGA -7.362 14.900 -0.494 0.6217   
## STABBRID -7.236 11.956 -0.605 0.5456   
## STABBRIL -31.170 12.139 -2.568 0.0108 \*   
## STABBRIN -15.278 14.469 -1.056 0.2921   
## STABBRIA 20.177 12.263 1.645 0.1012   
## STABBRKS 17.404 11.956 1.456 0.1468   
## STABBRKY -15.428 14.336 -1.076 0.2829   
## STABBRLA 45.860 19.831 2.313 0.0216 \*   
## STABBRME 2.513 12.865 0.195 0.8453   
## STABBRMI -14.302 10.567 -1.354 0.1771   
## STABBRMN -3.132 12.865 -0.243 0.8078   
## STABBRMS -13.089 14.900 -0.878 0.3805   
## STABBRMT 29.328 19.831 1.479 0.1405   
## STABBRNV 19.863 14.336 1.386 0.1672   
## STABBRNH -17.126 22.589 -0.758 0.4491   
## STABBRNJ -10.075 16.072 -0.627 0.5313   
## STABBRNM 37.463 23.886 1.568 0.1181   
## STABBRNY -1.927 8.024 -0.240 0.8104   
## STABBRNC 1.803 8.944 0.202 0.8404   
## STABBRND 29.320 19.831 1.479 0.1406   
## STABBROH -16.942 9.964 -1.700 0.0903 .   
## STABBROR -14.697 9.735 -1.510 0.1324   
## STABBRPA 3.095 8.695 0.356 0.7221   
## STABBRRI -28.465 15.275 -1.863 0.0636 .   
## STABBRSC -12.272 19.831 -0.619 0.5366   
## STABBRTN -31.531 12.711 -2.481 0.0138 \*   
## STABBRTX 13.941 14.336 0.972 0.3318   
## STABBRUT 26.065 11.956 2.180 0.0302 \*   
## STABBRVT -25.924 14.469 -1.792 0.0744 .   
## STABBRVA -5.876 9.605 -0.612 0.5413   
## STABBRWA -16.328 9.735 -1.677 0.0948 .   
## STABBRWY 53.291 19.831 2.687 0.0077 \*\*   
## Is.PartyTRUE -1.947 8.995 -0.216 0.8288   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.4 on 242 degrees of freedom  
## Multiple R-squared: 0.8272, Adjusted R-squared: 0.795   
## F-statistic: 25.73 on 45 and 242 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_4, model\_gender\_p6\_5)

## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender + CONTROL + School.Type + STABBR  
## Model 2: ROI ~ Gender + CONTROL + School.Type + STABBR + Is.Party  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 243 181757   
## 2 242 181722 1 35.178 0.0468 0.8288

## Tier is not helpful  
model\_gender\_p6\_6 = lm(ROI ~ Gender + CONTROL + School.Type + STABBR + Tier, data = data\_gender\_p6)  
summary(model\_gender\_p6\_6)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL + School.Type + STABBR +   
## Tier, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -81.405 -11.636 0.997 12.424 123.326   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 8.0303 10.7665 0.746  
## GenderMale 28.5130 3.2190 8.858  
## CONTROLPublic 103.0445 26.4078 3.902  
## School.TypeLiberal Arts -41.0548 8.0573 -5.095  
## School.TypeState -13.4165 13.8542 -0.968  
## School.TypeIvy League -31.1123 23.6540 -1.315  
## STABBRMA 1.8072 10.7329 0.168  
## STABBRAL -10.6397 14.3718 -0.740  
## STABBRAR 7.1174 19.8260 0.359  
## STABBRCO -0.9042 15.7267 -0.057  
## STABBRCT -12.6122 12.7837 -0.987  
## STABBRDE 13.6303 19.8260 0.687  
## STABBRFL 18.7860 9.6056 1.956  
## STABBRGA -6.6313 14.4430 -0.459  
## STABBRID -7.8276 12.0157 -0.651  
## STABBRIL -29.7344 12.3008 -2.417  
## STABBRIN -14.7373 14.7990 -0.996  
## STABBRIA 20.3152 12.1993 1.665  
## STABBRKS 16.8120 12.0157 1.399  
## STABBRKY -16.0203 14.3718 -1.115  
## STABBRLA 45.2680 19.8260 2.283  
## STABBRME 0.1385 12.9795 0.011  
## STABBRMI -14.8942 10.6438 -1.399  
## STABBRMN -1.5606 13.3140 -0.117  
## STABBRMS -14.6545 14.3718 -1.020  
## STABBRMT 28.7364 19.8260 1.449  
## STABBRNV 19.2716 14.3718 1.341  
## STABBRNH -15.9685 22.5320 -0.709  
## STABBRNJ -6.7510 16.2786 -0.415  
## STABBRNM 40.3134 24.0020 1.680  
## STABBRNY -0.6549 8.0973 -0.081  
## STABBRNC -2.8510 9.6604 -0.295  
## STABBRND 28.7279 19.8260 1.449  
## STABBROH -16.6096 10.0411 -1.654  
## STABBROR -15.1919 9.8043 -1.550  
## STABBRPA 3.0751 8.7519 0.351  
## STABBRRI -28.1824 15.2375 -1.850  
## STABBRSC -8.2713 20.5855 -0.402  
## STABBRTN -32.5058 12.9521 -2.510  
## STABBRTX 13.3494 14.3718 0.929  
## STABBRUT 25.4728 12.0157 2.120  
## STABBRVT -27.4068 14.4573 -1.896  
## STABBRVA -6.4347 9.5311 -0.675  
## STABBRWA -16.4495 9.7999 -1.679  
## STABBRWY 52.6993 19.8260 2.658  
## TierSelective public -18.5409 22.4037 -0.828  
## TierHighly selective private -1.8710 9.2414 -0.202  
## TierHighly selective public -23.1333 23.4699 -0.986  
## TierOther elite schools (public and private) 4.0482 7.6222 0.531  
## TierIvy Plus 41.1043 22.4032 1.835  
## Pr(>|t|)   
## (Intercept) 0.456488   
## GenderMale < 2e-16 \*\*\*  
## CONTROLPublic 0.000124 \*\*\*  
## School.TypeLiberal Arts 7.09e-07 \*\*\*  
## School.TypeState 0.333827   
## School.TypeIvy League 0.189672   
## STABBRMA 0.866431   
## STABBRAL 0.459839   
## STABBRAR 0.719918   
## STABBRCO 0.954198   
## STABBRCT 0.324846   
## STABBRDE 0.492438   
## STABBRFL 0.051666 .   
## STABBRGA 0.646557   
## STABBRID 0.515387   
## STABBRIL 0.016390 \*   
## STABBRIN 0.320342   
## STABBRIA 0.097174 .   
## STABBRKS 0.163064   
## STABBRKY 0.266102   
## STABBRLA 0.023297 \*   
## STABBRME 0.991494   
## STABBRMI 0.163013   
## STABBRMN 0.906789   
## STABBRMS 0.308920   
## STABBRMT 0.148534   
## STABBRNV 0.181222   
## STABBRNH 0.479202   
## STABBRNJ 0.678722   
## STABBRNM 0.094349 .   
## STABBRNY 0.935604   
## STABBRNC 0.768155   
## STABBRND 0.148654   
## STABBROH 0.099414 .   
## STABBROR 0.122586   
## STABBRPA 0.725627   
## STABBRRI 0.065620 .   
## STABBRSC 0.688190   
## STABBRTN 0.012748 \*   
## STABBRTX 0.353904   
## STABBRUT 0.035045 \*   
## STABBRVT 0.059211 .   
## STABBRVA 0.500247   
## STABBRWA 0.094554 .   
## STABBRWY 0.008391 \*\*   
## TierSelective public 0.408737   
## TierHighly selective private 0.839732   
## TierHighly selective public 0.325303   
## TierOther elite schools (public and private) 0.595838   
## TierIvy Plus 0.067791 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.31 on 238 degrees of freedom  
## Multiple R-squared: 0.8311, Adjusted R-squared: 0.7963   
## F-statistic: 23.9 on 49 and 238 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_4, model\_gender\_p6\_6)

## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender + CONTROL + School.Type + STABBR  
## Model 2: ROI ~ Gender + CONTROL + School.Type + STABBR + Tier  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 243 181757   
## 2 238 177567 5 4190.3 1.1233 0.3486

# Region is not helpful  
model\_gender\_p6\_7 = lm(ROI ~ Gender + CONTROL + School.Type + STABBR + Region, data = data\_gender\_p6)  
summary(model\_gender\_p6\_7)

##   
## Call:  
## lm(formula = ROI ~ Gender + CONTROL + School.Type + STABBR +   
## Region, data = data\_gender\_p6)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.723 -13.174 0.123 12.095 124.008   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.585 8.928 1.298 0.19567   
## GenderMale 28.513 3.223 8.846 < 2e-16 \*\*\*  
## CONTROLPublic 79.272 14.630 5.418 1.45e-07 \*\*\*  
## School.TypeLiberal Arts -43.020 7.760 -5.544 7.70e-08 \*\*\*  
## School.TypeState -12.422 13.451 -0.923 0.35668   
## School.TypeIvy League 7.513 11.903 0.631 0.52851   
## STABBRMA 7.798 10.288 0.758 0.44920   
## STABBRAL -9.958 14.302 -0.696 0.48693   
## STABBRAR 7.799 19.787 0.394 0.69381   
## STABBRCO -1.816 15.339 -0.118 0.90588   
## STABBRCT -13.454 12.608 -1.067 0.28697   
## STABBRDE 14.312 19.787 0.723 0.47018   
## STABBRFL 18.549 9.610 1.930 0.05473 .   
## STABBRGA -8.245 14.302 -0.577 0.56479   
## STABBRID -7.146 11.925 -0.599 0.54960   
## STABBRIL -31.088 12.110 -2.567 0.01085 \*   
## STABBRIN -15.191 14.435 -1.052 0.29368   
## STABBRIA 19.617 11.963 1.640 0.10235   
## STABBRKS 17.494 11.925 1.467 0.14368   
## STABBRKY -15.338 14.302 -1.072 0.28457   
## STABBRLA 45.950 19.787 2.322 0.02105 \*   
## STABBRME 2.598 12.833 0.202 0.83977   
## STABBRMI -14.212 10.538 -1.349 0.17869   
## STABBRMN -3.048 12.833 -0.237 0.81249   
## STABBRMS -13.973 14.302 -0.977 0.32955   
## STABBRMT 29.418 19.787 1.487 0.13838   
## STABBRNV 19.954 14.302 1.395 0.16424   
## STABBRNH -17.044 22.541 -0.756 0.45031   
## STABBRNJ -10.001 16.037 -0.624 0.53345   
## STABBRNM 37.398 23.837 1.569 0.11797   
## STABBRNY -1.845 8.000 -0.231 0.81781   
## STABBRNC 1.892 8.917 0.212 0.83215   
## STABBRND 29.410 19.787 1.486 0.13849   
## STABBROH -16.855 9.936 -1.696 0.09110 .   
## STABBROR -14.609 9.707 -1.505 0.13363   
## STABBRPA 3.178 8.669 0.367 0.71426   
## STABBRRI -28.379 15.240 -1.862 0.06380 .   
## STABBRSC -12.182 19.787 -0.616 0.53871   
## STABBRTN -31.492 12.684 -2.483 0.01371 \*   
## STABBRTX 14.031 14.302 0.981 0.32752   
## STABBRUT 26.155 11.925 2.193 0.02924 \*   
## STABBRVT -25.836 14.435 -1.790 0.07473 .   
## STABBRVA -6.139 9.509 -0.646 0.51919   
## STABBRWA -16.241 9.707 -1.673 0.09560 .   
## STABBRWY 53.381 19.787 2.698 0.00747 \*\*   
## RegionNortheastern NA NA NA NA   
## RegionSouthern NA NA NA NA   
## RegionWestern NA NA NA NA   
## RegionMidwestern NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27.35 on 243 degrees of freedom  
## Multiple R-squared: 0.8271, Adjusted R-squared: 0.7958   
## F-statistic: 26.42 on 44 and 243 DF, p-value: < 2.2e-16

anova(model\_gender\_p6\_5, model\_gender\_p6\_7)

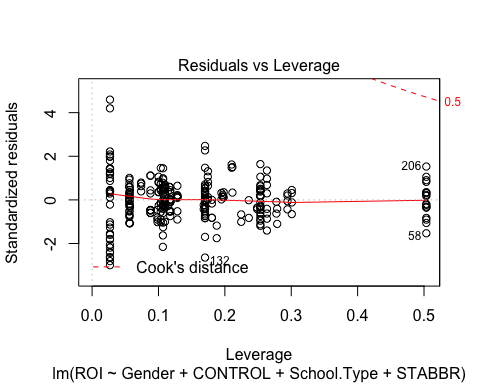
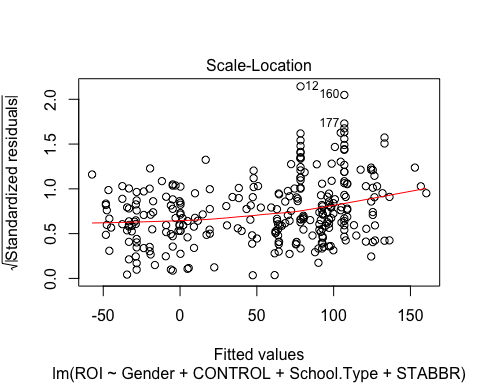
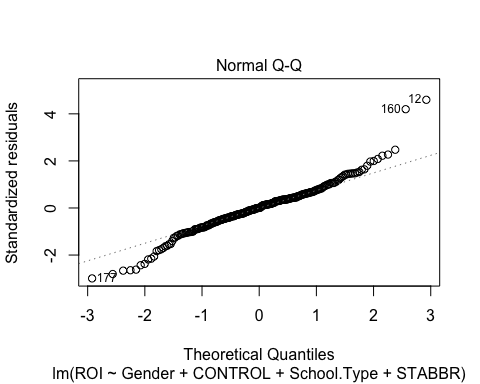
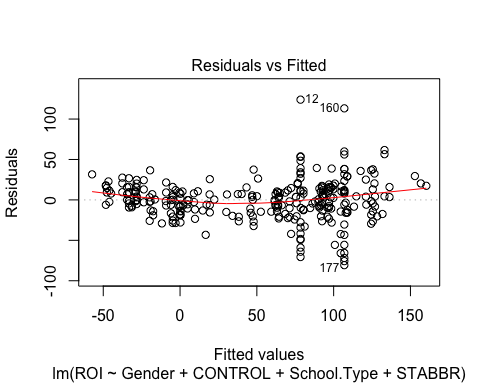
## Analysis of Variance Table  
##   
## Model 1: ROI ~ Gender + CONTROL + School.Type + STABBR + Is.Party  
## Model 2: ROI ~ Gender + CONTROL + School.Type + STABBR + Region  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 242 181722   
## 2 243 181757 -1 -35.178 0.0468 0.8288

# Final model: ROI ~ Gender + CONTROL + School.Type + STABBR

By using hierarchical regression, we find the best model to predict starting salary based on gender is ROI ~ Gender + CONTROL + School.Type + STABBR. Tier, Is.Party and Region will not help to improve the model.

The model meets the assumptions of linear regression:

plot(model\_gender\_p6\_4)



Here is the table for coefficients:

kable(coefficients(summary(model\_gender\_p6\_4)))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 11.584811 | 8.928113 | 1.2975655 | 0.1956671 |
| GenderMale | 28.512982 | 3.223119 | 8.8463936 | 0.0000000 |
| CONTROLPublic | 79.272481 | 14.630440 | 5.4183253 | 0.0000001 |
| School.TypeLiberal Arts | -43.020185 | 7.760014 | -5.5438283 | 0.0000001 |
| School.TypeState | -12.421786 | 13.451167 | -0.9234728 | 0.3566769 |
| School.TypeIvy League | 7.512809 | 11.902573 | 0.6311920 | 0.5285082 |
| STABBRMA | 7.798325 | 10.288391 | 0.7579732 | 0.4492014 |
| STABBRAL | -9.957687 | 14.301788 | -0.6962547 | 0.4869348 |
| STABBRAR | 7.799404 | 19.787221 | 0.3941637 | 0.6938059 |
| STABBRCO | -1.815523 | 15.338615 | -0.1183629 | 0.9058779 |
| STABBRCT | -13.454461 | 12.608083 | -1.0671298 | 0.2869723 |
| STABBRDE | 14.312327 | 19.787221 | 0.7233116 | 0.4701841 |
| STABBRFL | 18.549463 | 9.609649 | 1.9302955 | 0.0547340 |
| STABBRGA | -8.245492 | 14.301788 | -0.5765357 | 0.5647868 |
| STABBRID | -7.145573 | 11.925190 | -0.5991999 | 0.5495978 |
| STABBRIL | -31.088280 | 12.109585 | -2.5672457 | 0.0108506 |
| STABBRIN | -15.190913 | 14.435175 | -1.0523540 | 0.2936825 |
| STABBRIA | 19.616519 | 11.963229 | 1.6397344 | 0.1023543 |
| STABBRKS | 17.493960 | 11.925190 | 1.4669753 | 0.1436758 |
| STABBRKY | -15.338321 | 14.301788 | -1.0724758 | 0.2845703 |
| STABBRLA | 45.950009 | 19.787221 | 2.3222063 | 0.0210490 |
| STABBRME | 2.597595 | 12.833425 | 0.2024085 | 0.8397667 |
| STABBRMI | -14.212255 | 10.537773 | -1.3486962 | 0.1786901 |
| STABBRMN | -3.047658 | 12.833425 | -0.2374781 | 0.8124860 |
| STABBRMS | -13.972539 | 14.301788 | -0.9769785 | 0.3295514 |
| STABBRMT | 29.418366 | 19.787221 | 1.4867356 | 0.1383810 |
| STABBRNV | 19.953568 | 14.301788 | 1.3951799 | 0.1642354 |
| STABBRNH | -17.043803 | 22.541267 | -0.7561156 | 0.4503123 |
| STABBRNJ | -10.001437 | 16.037256 | -0.6236377 | 0.5334505 |
| STABBRNM | 37.397676 | 23.837002 | 1.5688918 | 0.1179744 |
| STABBRNY | -1.844846 | 7.999590 | -0.2306176 | 0.8178060 |
| STABBRNC | 1.891970 | 8.916847 | 0.2121793 | 0.8321450 |
| STABBRND | 29.409845 | 19.787221 | 1.4863050 | 0.1384947 |
| STABBROH | -16.854879 | 9.935973 | -1.6963491 | 0.0911004 |
| STABBROR | -14.608721 | 9.707019 | -1.5049646 | 0.1336319 |
| STABBRPA | 3.177807 | 8.669176 | 0.3665639 | 0.7142633 |
| STABBRRI | -28.379062 | 15.240329 | -1.8621030 | 0.0637956 |
| STABBRSC | -12.181734 | 19.787221 | -0.6156364 | 0.5387107 |
| STABBRTN | -31.492290 | 12.684326 | -2.4827720 | 0.0137125 |
| STABBRTX | 14.031373 | 14.301788 | 0.9810922 | 0.3275231 |
| STABBRUT | 26.154747 | 11.925190 | 2.1932353 | 0.0292385 |
| STABBRVT | -25.836236 | 14.435175 | -1.7898111 | 0.0747294 |
| STABBRVA | -6.138540 | 9.509216 | -0.6455359 | 0.5191890 |
| STABBRWA | -16.240519 | 9.707019 | -1.6730696 | 0.0956004 |
| STABBRWY | 53.381302 | 19.787221 | 2.6977665 | 0.0074698 |

Overall, this model can explain about 80% of the variance (Adjusted R-squared=0.80) and very significant (p=.00).

To be specific:

* Gender: Compared to female students, male students will significantly(p=.00) improve their ROI about 28.51%.
* Tier: Tier (Selectivity) information is partially correlated with CONTROL because some of them are dedicated public or private category. It also correlate with School.Type, for example Ivy League category. And the ANOVA test of model shows that it will not help to improve model when both CONTROL and School.Type are present.
* Control: Whether a school is public or private weight most in the model. If a student go to a public school, his/her ROI will go up about 79.27% compared to a student go to a private school. This probably caused by the lower cost of public schools.
* School.Type: Compared to students go to engineering school, students who go to Liberal Arts school will have much lower ROI by 43.02%. This probably caused by the low starting salary for those students.
* STABBR: Compared to students in California, students in Wyoming, Louisiana and Utah will get higher ROI, but students in Illinois, Tennessee will get lower ROI.

We can’t do mediation test since all independent variables are not numerical. We can’t compute ‘a’ path model.