440 Reproducibility and Statistics Assignment Homework

```
politics<-read.csv("politics.csv")
```

I loaded the politics.csv data file.

```
str(politics)
```

```
## 'data.frame':    132 obs. of  7 variables:
##  $ subject      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ party        : Factor w/ 3 levels "democrat","independent",..: 3 3 2 2 2 3 3 2 3 2 ...
##  $ testtime     : Factor w/ 2 levels "post","pre": 2 2 2 2 2 2 2 2 2 2 ...
##  $ optimismscore: int  52 51 69 51 61 31 57 48 42 64 ...
##  $ minwage      : Factor w/ 2 levels "no","yes": 1 1 2 1 2 1 1 1 1 1 ...
##  $ sex          : Factor w/ 2 levels "female","male": 2 2 1 2 2 2 2 2 2 2 ...
##  $ income       : num  37.3 42.3 73 33.8 57.3 ...
```

I looked at the data.

```
politics$subject<-factor(politics$subject)
```

I changed subject into a factor variable.

```
politics$testtime<-factor(politics$testtime, levels=c("pre", "post"))
```

I refactored the variables “pre” and “post” so that “pre” precedes “post”.

```
summary(politics$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.672  27.590  41.220  43.040  56.010 114.800
```

I found the minimum, mean, and median incomes, but I want the minimun, mean, and variance for posttest optimism scores.

```
summary(politics$optimismscore[politics$testtime=="post"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   45.00   61.00   59.82   73.00   94.00
```

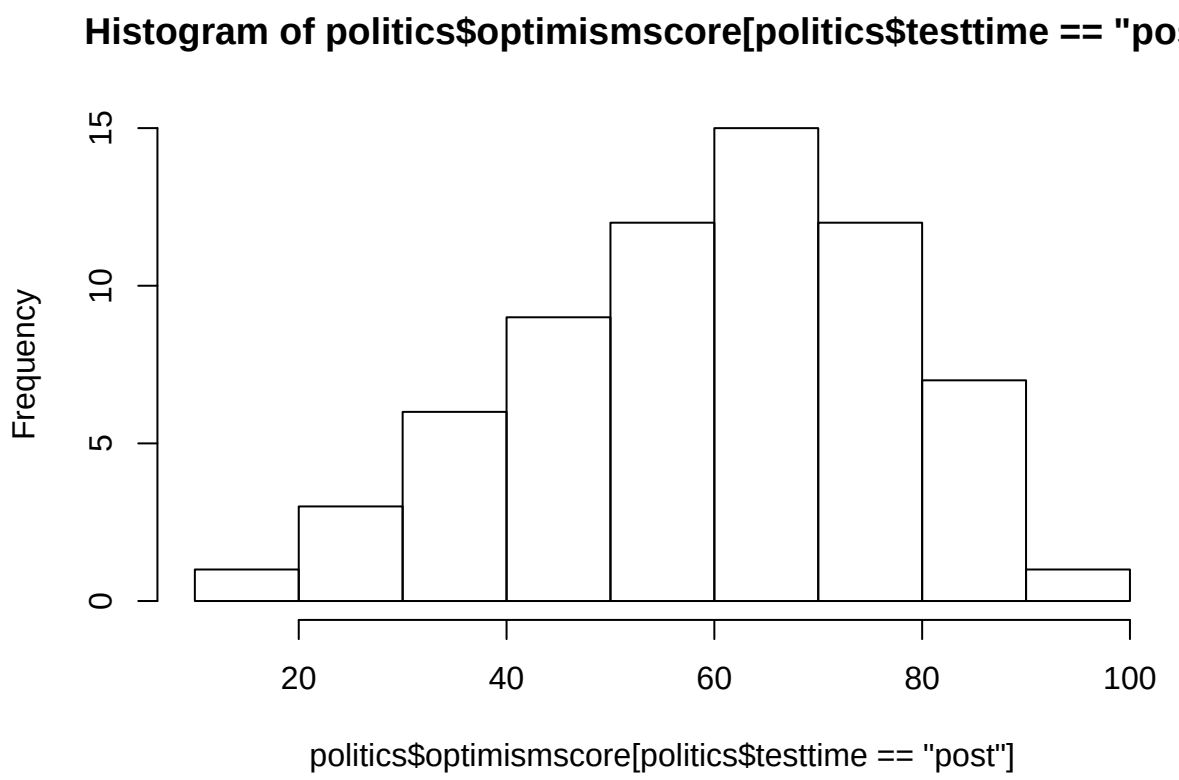
I found the minimum, mean, and median for posttest optimisim scores.

```
var(politics$optimismscore[politics$testtime=="post"])
```

```
## [1] 336.2741
```

I found the variance for posttest optimism scores.

```r
hist(politics$optimismscore[politics$testtime=="post"])
```



I created a histogram ofposttest optimism scores.

```r
tab<-table(politics$party[politics$testtime=="post"],politics$sex[politics$testtime=="post"])
tab
```

```
##
##               female male
##   democrat        14   12
##   independent      7   10
##   republican      12   11
```

I used a table to calulate the frequency that individuals appear in different political groups. I focused on the posttest data to avoid counting people twice.

```r
chisq.test(politics$party[politics$testtime=="post"],politics$sex[politics$testtime=="post"])
```

```
##
##  Pearson's Chi-squared test
##
## data:  politics$party[politics$testtime == "post"] and politics$sex[politics$testtime == "post"]
## X-squared = 0.7267, df = 2, p-value = 0.6953
```

I used a Chi-Square test of independence to test the hypothesis that affiliation and support are independent. I focused on the posttest data to avoid counting people twice.Political affiliation is independent of gender, Chi-Square [2]= 0.73, p-value = 0.70.

```
t.test(politics$income[politics$sex=="male" & politics$testtime=="post"],politics$income[politics$sex==
```

```
##
##  Welch Two Sample t-test
##
## data:  politics$income[politics$sex == "male" & politics$testtime ==  and politics$income[politics$s
## t = -1.5714, df = 61.623, p-value = 0.1212
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.23627   2.30508
## sample estimates:
## mean of x mean of y
##  38.80751  47.27310
```

To determine whether or not males and females have different posttest incomes, I ran an independent t-test. Males and females have different incomes, t(61.6) = -1.5714, p-value = 0.12

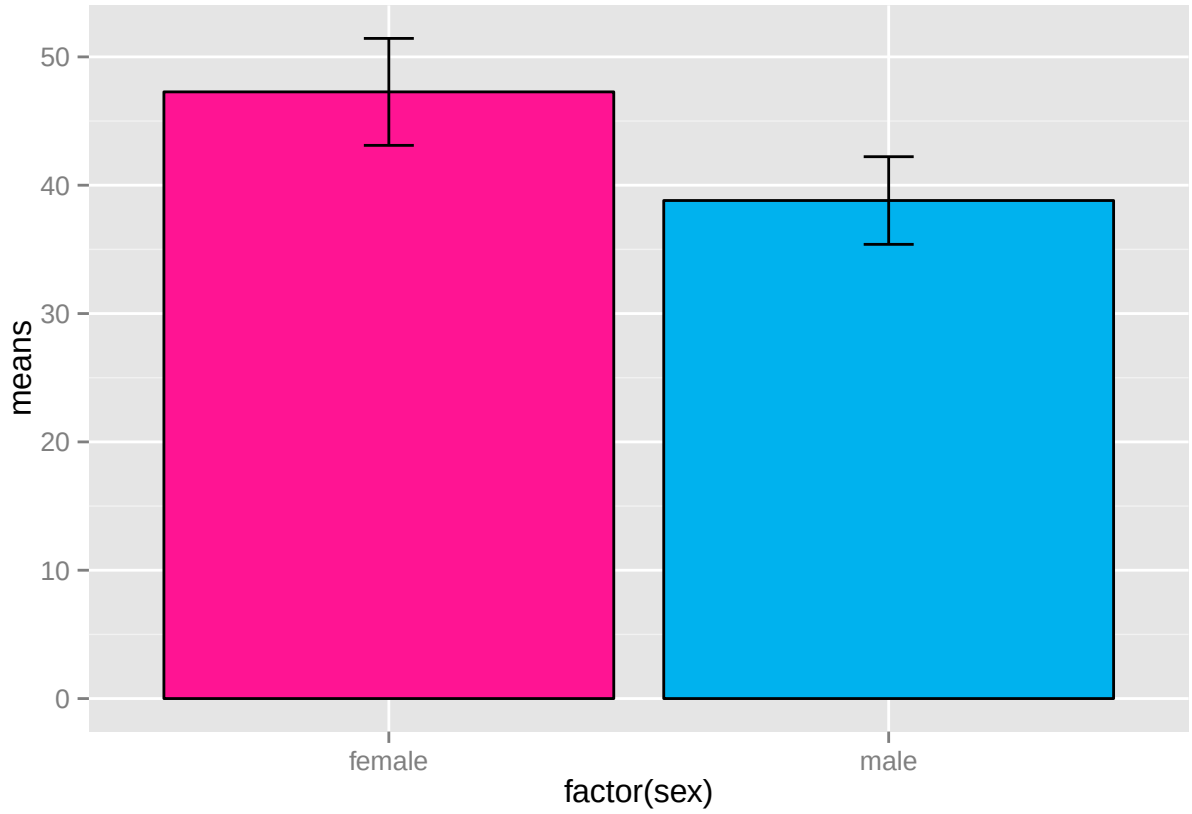
```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
temp<-politics[politics$testtime=="post",]%>%group_by(sex)%>%summarize(means=mean(income),
              sems=sd(income)/sqrt(length(income)))
library("ggplot2")
```

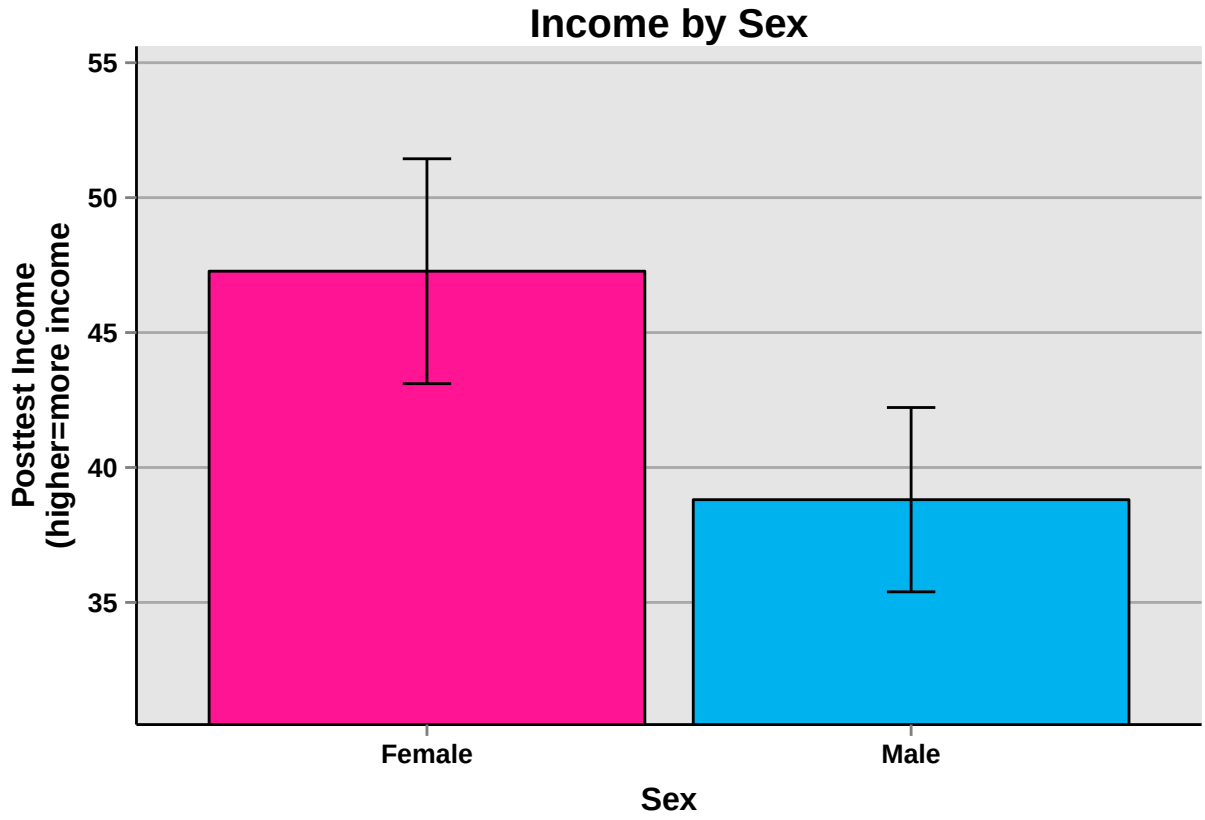
```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
f<-ggplot(temp,aes(x=factor(sex),y=means))+
    geom_bar(stat="identity", color="black",fill=c("deeppink","deepskyblue2"))+
    geom_errorbar(aes(ymax=means+sems, ymin=means-sems), width=.1)
f
```



I created a figure by using the dplyr library.

```
f<-f+ggtitle("Income by Sex")+
    labs(x="Sex", y="Posttest Income\n(higher=more income")+
    scale_x_discrete(breaks=c("female","male"),labels=c("Female","Male"))+
    theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
    theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
    theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
    theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
    theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
    coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),max(temp$means)+2*max(temp$sems)))+
    theme(panel.border=element_blank(), axis.line=element_line())+
    theme(panel.grid.major.x=element_blank())+
    theme(panel.grid.major.y=element_line(color="darkgrey"))+
    theme(panel.grid.minor.y=element_blank())
f
```



I made the graph more aesthetically pleasing and informative by changing the axis labels and text.

```
summary(aov(optimismscore~party*sex,data=politics[politics$testtime=="post",]))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## party        2  10147    5074  27.063 4.2e-09 ***
## sex          1      7       7   0.040   0.843
## party:sex    2    455     227   1.213   0.304
## Residuals   60  11248     187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I used a 2-way between-subjects ANOVA to seef if party affiliation and sex predict posttest optimism scores independently or in an interaction. The p-values for sex 0.843 and party:sex 0.304 do not tell me much in particular. Maybe there's nothing going on, or maybe I just can't see it. The super tiny P- value for Party 4.2e-09 is definitely something. The two party columns have meaningfully different averages.

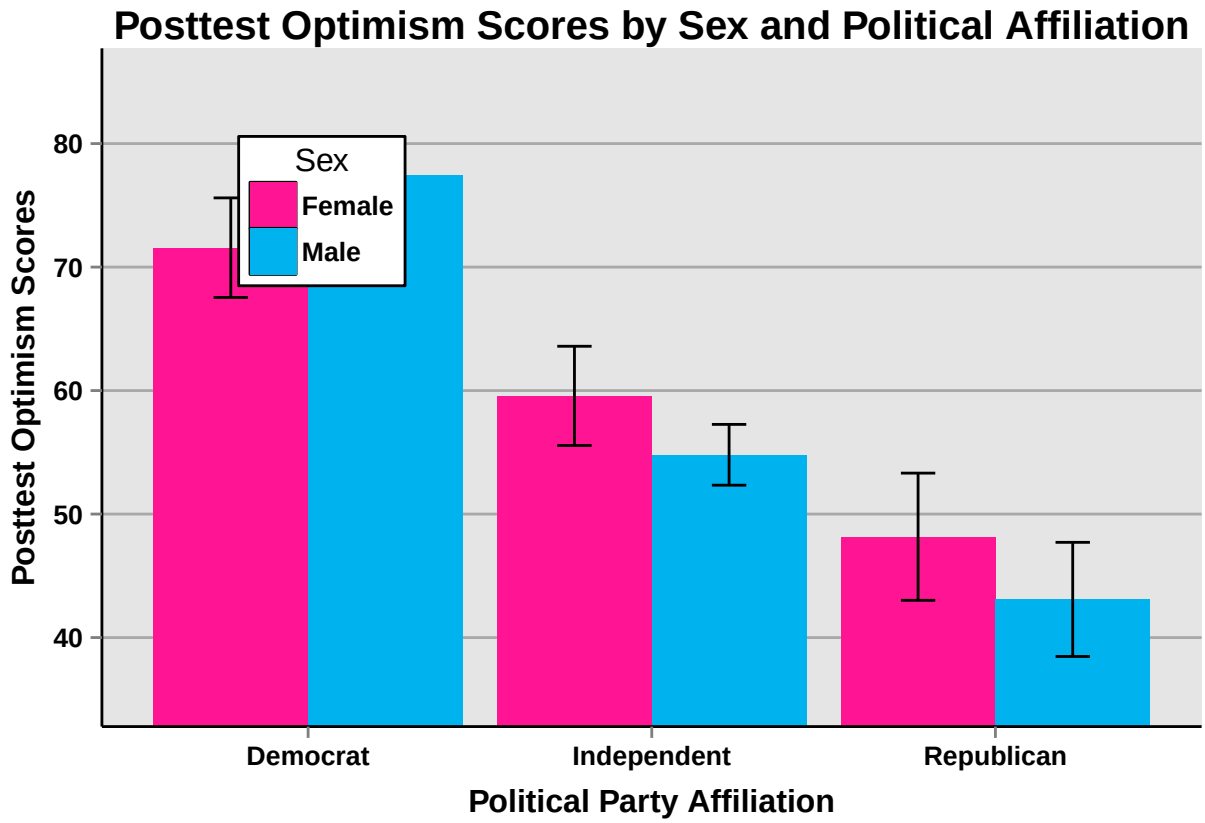
```
temp<-politics[politics$testtime=="post",]%>%group_by(party,sex)%>%
    summarize(means=mean(optimismscore),sems=sd(optimismscore)/sqrt(length(optimismscore)))
library("gplots")
```

```
## Warning: package 'gplots' was built under R version 3.1.3
```

```
##
## Attaching package: 'gplots'
```

```
##
## The following object is masked from 'package:stats':
##
##     lowess
```

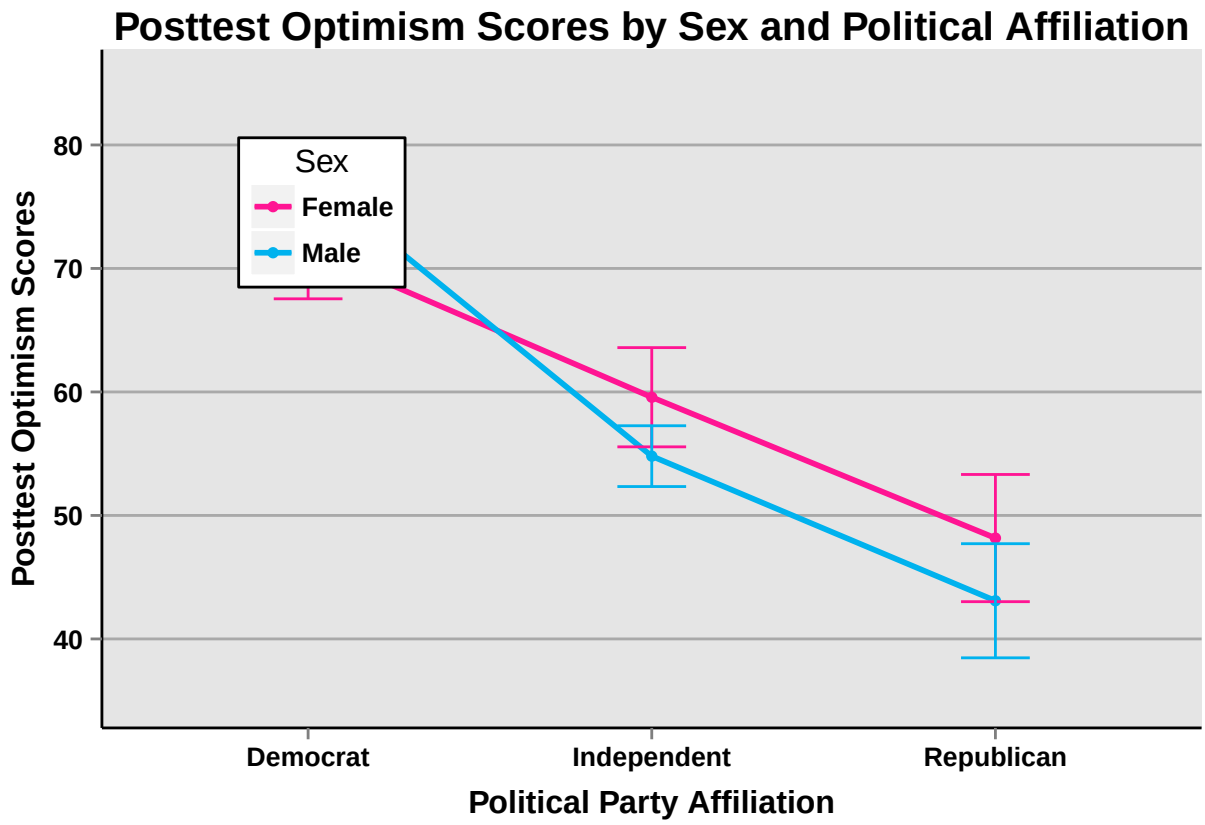
```
col1=col2hex("deeppink")
col2=col2hex("deepskyblue2")
f<-ggplot(temp, aes(x=party, y=means, fill=sex))+
    geom_bar(stat="identity",position=position_dodge())+
    scale_fill_manual(values=c(col1,col2),name="Sex",breaks=c("female","male"),labels=c("Female", "Male
    theme(legend.key=element_rect(color="black"))+
    geom_errorbar(aes(ymax=means+sems, ymin=means-sems),width=.2,position=position_dodge(.9))+
    ggtitle("Posttest Optimism Scores by Sex and Political Affiliation")+
    labs(x="Political Party Affiliation",y="Posttest Optimism Scores")+
    scale_x_discrete(breaks=c("democrat","independent","republican"),labels=c("Democrat","Independent",
    theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
    theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
    theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
    theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
    theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
    coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),max(temp$means)+2*max(temp$sems)))+
    theme(panel.border=element_blank(),axis.line=element_line())+
    theme(panel.grid.major.x=element_blank())+
    theme(panel.grid.major.y=element_line(color="darkgrey"))+
    theme(panel.grid.minor.y=element_blank())+
    theme(legend.position=c(.2,.76))+
    theme(legend.background=element_blank())+
    theme(legend.background=element_rect(color="black"))+
    theme(legend.title=element_blank())+
    theme(legend.title=element_text(size=12))+
    theme(legend.title.align=.5)+
    theme(legend.text=element_text(size=10,face="bold"))
f
```



I created a plot via gplots library.

```
f<-ggplot(temp, aes(x=party, y=means, group=sex, color=sex))+
    geom_line(size=1)+
    geom_point(size=2)+
    scale_color_manual(values=c(col1,col2),name="Sex",breaks=c("female","male"),labels=c("Female", "Mal
    geom_errorbar(aes(ymax=means+sems, ymin=means-sems),width=.2)+
    ggtitle("Posttest Optimism Scores by Sex and Political Affiliation")+
    labs(x="Political Party Affiliation",y="Posttest Optimism Scores")+
    scale_x_discrete(breaks=c("democrat","independent","republican"),labels=c("Democrat","Independent",
    theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
    theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
    theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
    theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
    theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
    coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),max(temp$means)+2*max(temp$sems)))+
    theme(panel.border=element_blank(),axis.line=element_line())+
    theme(panel.grid.major.x=element_blank())+
    theme(panel.grid.major.y=element_line(color="darkgrey"))+
    theme(panel.grid.minor.y=element_blank())+
    theme(legend.position=c(.2,.76))+
    theme(legend.background=element_blank())+
    theme(legend.background=element_rect(color="black"))+
    theme(legend.title=element_blank())+
    theme(legend.title=element_text(size=12))+
    theme(legend.title.align=.5)+
```

```
    theme(legend.text=element_text(size=10,face="bold"))
f
```



I created a line graph for the same data with error bars.

```
summary(aov(optimismscore~testtime*sex+Error(subject/testtime),data=politics))
```

```
##
## Error: subject
##           Df Sum Sq Mean Sq F value Pr(>F)
## sex        1     80    80.4   0.119  0.731
## Residuals 64  43105   673.5
##
## Error: subject:testtime
##              Df Sum Sq Mean Sq F value   Pr(>F)
## testtime      1  770.9   770.9  41.299 1.87e-08 ***
## testtime:sex  1    0.9     0.9   0.049    0.825
## Residuals    64 1194.7    18.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To see if optimism scores before and after watching videos vary depending on sex, I ran a 2-way mixed ANOVA. I used a mixed ANOVA because the same people indicated their optimism before and after the videos (i.e. a within-subjects factor) and because there are different genders affiliated with different optimism scores (i.e.a between-subjects factor). The testtime p-value is 1.87e-08. Since the testtime groups have different averages, it indicates that optimism scores varied.

```
summary(lm(optimismscore~politics$optimismscore[politics$testtime=="pre"]+party,data=politics[politics$
```

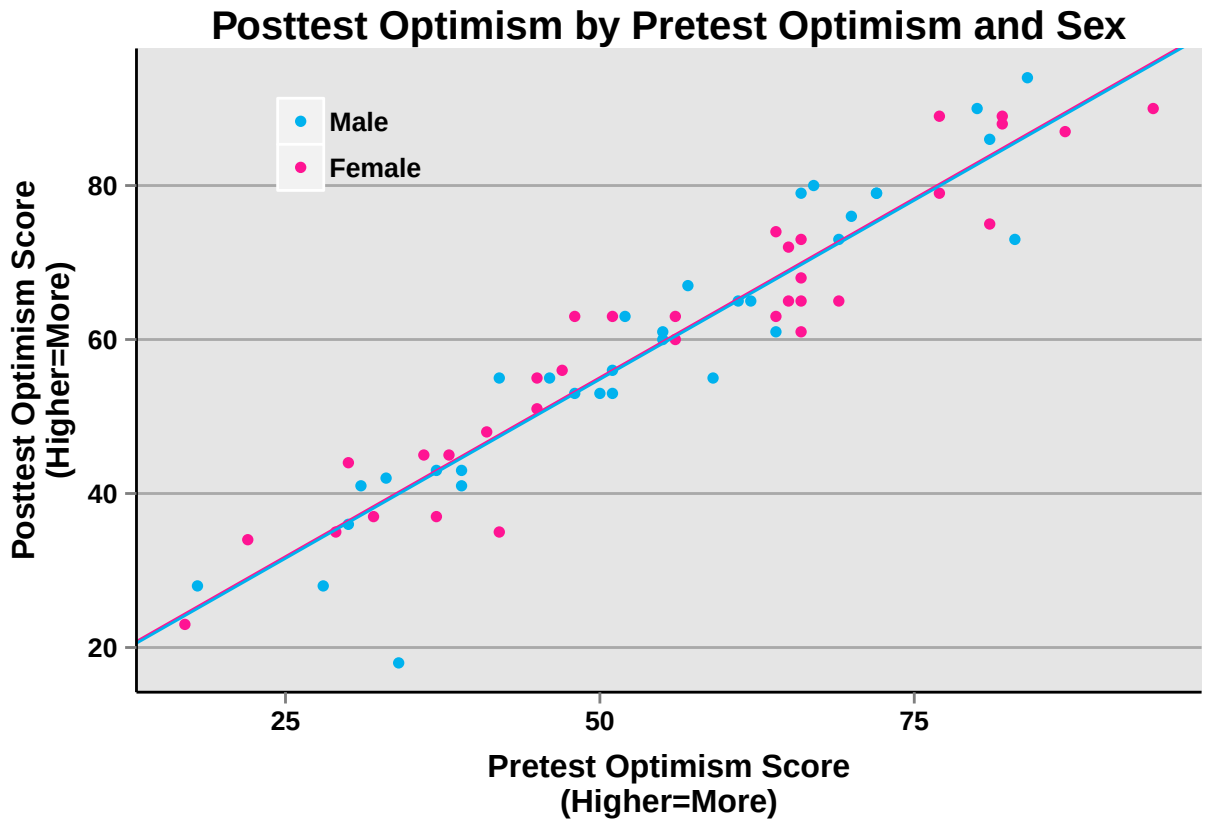
```
##
## Call:
## lm(formula = optimismscore ~ politics$optimismscore[politics$testtime ==
##     "pre"] + party, data = politics[politics$testtime == "post",
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.866  -2.562   1.267   3.901   8.948
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                          8.1219     4.2616
## politics$optimismscore[politics$testtime == "pre"]   0.9419     0.0583
## partyindependent                                    -1.3310     2.1284
## partyrepublican                                      0.7210     2.5000
##                                                    t value Pr(>|t|)
## (Intercept)                                          1.906   0.0613 .
## politics$optimismscore[politics$testtime == "pre"]  16.154   <2e-16 ***
## partyindependent                                    -0.625   0.5340
## partyrepublican                                      0.288   0.7740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.022 on 62 degrees of freedom
## Multiple R-squared:  0.8971, Adjusted R-squared:  0.8922
## F-statistic: 180.3 on 3 and 62 DF,  p-value: < 2.2e-16
```

I used Multiple Regression to predict posttime optimism scores from pretest optimism scores and party affiliation. The Multiple R-squared=0.90, F-statistic=180.3 on 3 and 62 DF, and the p-value= 2.2e-16 indicate a lack of significance.

```
#summary(lm(optimismscore~OptScorePre+sex,data=politics[politics$testtime=="post",]))No OptScorePre so
```

```
OptScorePre<-politics$optimismscore[politics$testtime=="pre"]
f<-ggplot(politics[politics$testtime=="post",],aes(x=OptScorePre,y=optimismscore,color=sex))+
    geom_point(size=2)+
    geom_abline(intercept=8.44+0.22/2, slope=0.93,color=col1)+
    geom_abline(intercept=8.44-0.22/2, slope=0.93,color=col2)+
    scale_color_manual(values=c(col1,col2),breaks=c("male","female"),labels=c("Male","Female"))+
    ggtitle("Posttest Optimism by Pretest Optimism and Sex")+
    labs(x="Pretest Optimism Score\n(Higher=More)",y="Posttest Optimism Score\n(Higher=More)")+
    theme(plot.title=element_text(size=15,face="bold", vjust=.5))+
    theme(axis.title.x=element_text(size=12,face="bold", vjust=-.25))+
    theme(axis.title.y=element_text(size=12,face="bold", vjust=1))+
    theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
    theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
    theme(panel.border=element_blank(), axis.line=element_line())+
    theme(panel.grid.major.x=element_blank())+
    theme(panel.grid.minor.x=element_blank())+
```

```
    theme(panel.grid.major.y=element_line(color="darkgrey"))+
    theme(panel.grid.minor.y=element_blank())+
    theme(legend.position=c(.2,.86))+
    theme(legend.background=element_blank())+
    theme(legend.title=element_blank())+
    theme(legend.text=element_text(size=10,face="bold"))
f
```



I ran a regression using sex instead of party, so that I can find intercepts and create the plot as instructed. I created a scatter plot. The Multiple R-squared=0.90, F-statistic=269.6 on 2 and 63 DF, and the p-value= 2.2e-16 indicate a lack of significance.

Fin