Final Data Analysis Project
Candice Yao(N11895239)

The key dependent variable of interest(Y)in this project will be **satsoc, "R's satisfaction with social activities and relationships"**. This project will study the effects caused by other variables on one's **satsoc**.

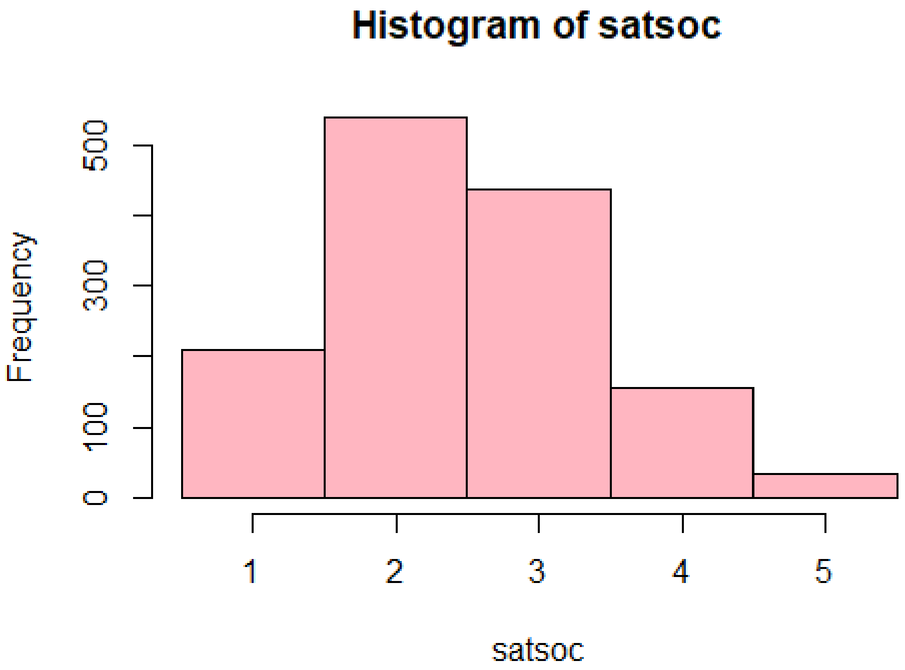
(a)histogram, sample mean and sample standard deviation

Code used:

```
hist((satsoc),
prob = TRUE,
main ="Histogram of social satisfaction",
xlab="social satisfaction",
col="lightpink")
hist(satsoc, breaks = seq(from=0.5, to=5.5,
by=1 ),col="lightpink")

mean(data$satsoc)

sd(data$satsoc)
```



The histogram that visualizes the distribution of 5 different scores of satsoc is plotted as above. The sample mean of satsoc is 2.466 and the sample standard deviation is

0.964.

(b) 95% confidence interval of Y

As calculated above, sd= 0.964, mean = 2.466

Codes used:

```
n= nrow(data)
margin_error = 1.96 *sd/sqrt(n)

lower_bound = mean - margin_error

upper_bound = mean + margin_error
```

Therefore, therefore, margin_error = 0.051, the lower bound of the confidence interval is 2.415 and the upper bound is 2.517.

Summing up, the 95% confidence interval of one's social satisfaction is between 2.415 and 2.517, which means that we are 95% percent sure that the true mean lies within this interval.

(c)Hypothesis test of the difference in the mean of Y between male and female at 5% significance level.

In this two-sample t test for social satisfaction by respondent’s sex. The null hypothesis(H0) is that there is no difference between male and female’s social satisfaction, while the alternative hypothesis(H1) is that there is a significant difference between male and female's social satisfaction.

Codes used:

```
mean_men = mean(satsoc[sex==1])
mean_women= mean(satsoc[sex==2])
t.test(satsoc[sex==1],satsoc[sex==2])
```

Test Results

```
Welch Two Sample t-test data:
satsoc[sex == 1] and satsoc[sex == 2]
t = -0.092868, df = 1353.7, p-value = 0.926
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-0.10710764  0.09742506
sample estimates:
mean of x mean of y
2.463303  2.468144
```

Since the p value is really large, exceeding the alpha(0.926>0.05), and the 95%

confidence interval for the difference in sample means(-0.107 to 0.097) includes 0, we don't have statistically significant evidence to reject the null hypothesis. Therefore, we don’t have sufficient evidence to claim that the difference in a respondent’s sex has a significant effect on a respondent’s satsoc.

(d)Hypothesizing that a respondent’s quality of life could affect their satsoc, I will construct a binary regression model between X(quallife, “R's quality of life”) and Y(satsoc, “R's satisfaction with social activities and relationships”)

Codes used:

```
model2 <- lm(satsoc ~ quallife, data=data)
summary(model2)
```

Regression results:

```
Call:
lm(formula = satsoc ~ quallife, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4091 -0.4091  0.1119  0.6330  2.6330 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.32487    0.05948   22.27   <2e-16 ***
quallife     0.52107    0.02511   20.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.8417 on 1374 degrees of freedom
Multiple R-squared:  0.2386,	Adjusted R-squared:  0.2381 
F-statistic: 430.6 on 1 and 1374 DF,  p-value: < 2.2e-16
```

According to the results above, the slope of the equation between quallife and satsoc is 0.521. The Adjusted R-squared is 0.2381.

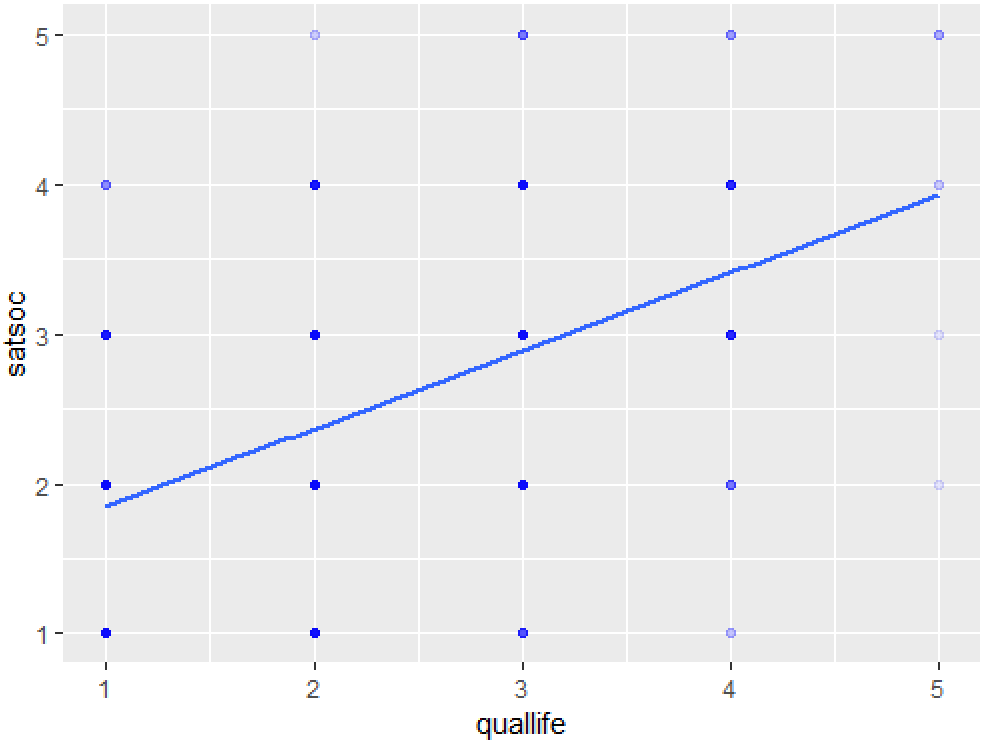
Since the slope is 0.521, we can tell that the increase in each unit of R’s quallife will result in 0.521 units of increase in one's social satisfaction. Since P value is relatively small with a high significance level, we can infer that this linear relationship is statistically significant.

Besides, the R-squared value is 0.2381, indicating that it 23.8% of the variance in Y(satsoc) can be explained by variance in X(quallife).

(e).Scatter plot of X(quallife, “R's quality of life”) and Y(satsoc, one's satisfaction with social activities and relationships)

Codes used:

```
ggplot(data)  +aes(x  =  quallife,  y  =satsoc  )  +
geom_point(alpha= .05, color = "blue" )
cor(data$age, data$satsoc,use = "complete.obs" )
ggplot(data)  +aes(x  =  quallife,  y  =satsoc  )  +
geom_point(alpha= .05, color = "blue" ) +geom_smooth(method
= lm, se = FALSE)
```



(f)Multiple regression model

Other factors that might determine satsoc might include R’s physical health and R’s race. I will conduct a multiple regression model using **hlthphys(R’s physical health)**, **hlthmntl(R’s mental health, mood, and ability to think)** and **race(race of respondent)** to predict **satsoc**.

Since the subject of investigation is concerning one’s satisfaction with social activities and relationships, a repsondent’s physical health might have influence over satsoc since disabilities could limit one’s access to participate in social events or mentally discourage people to socialize. Ones’ mental health or thinking capability could influence satsoc since many mental illness could result in self-isolation and this could negatively affect one’s social life. Lastly, one’s race might possibly influence one’s satsoc, since systematic discrimination and social pressure could be crucial factors concerning whether a non-white respondent could establish satisfactory social relationship.

Codes used:

```
model3  <-  lm(satsoc  ~   hlthphys  +  as.factor(race)  +
hlthmntl,data = data)
summary(model3)
```

Regression results:

```
Call:
lm(formula = satsoc ~ hlthphys + as.factor(race) + hlthmntl,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6073 -0.4848 -0.1276  0.4718  2.6026 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.92604    0.06259  14.796  < 2e-16 ***
hlthphys          0.13076    0.02259   5.788 8.82e-09 ***
as.factor(race)2  0.01778    0.06601   0.269    0.788    
as.factor(race)3 -0.12561    0.06431  -1.953    0.051 .  
hlthmntl          0.53955    0.02507  21.519  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.7654 on 1371 degrees of freedom
Multiple R-squared:  0.3718,  Adjusted R-squared:   0.37 
F-statistic: 202.8 on 4 and 1371 DF,  p-value: < 2.2e-16
```

Coefficients:

Factor 1: From the multiple regression model above, the coefficient of physical health is 0.131 with a p-value of 8.82e-09. Therefore, holding the other two factors constant, the increase of every one unit in a respondent’s physical health will lead to increase of 0.131 units in their satsoc. This means that when other factors are controlled, the higher physical health score(ranges from 1-5) one gets, the higher social satisfaction score one might obtain(ranges from 1-5), and vice versa. It is noteworthy that since the scores are coded inversely, the lower one’s score, the more healthy/more socially satisfactory one is. Therefore, the data could be understood as that the more physically unhealthy one is, one will be more unsatisfactory with their social activities and relationships, and vice versa. The result is statistically significant at a 95% confidence interval since the p value is extremely small.

Factor 2: Race in this data set is a categorical data, so it is coded as three different factors, and (race)1,white, is the reference group. The coefficient of (race)2, black, is

0.018, meaning that holding the other two factors constant, compared to being white, the reference group, being black will result in an increase of 0.018 units in satsoc. However, there is no evidence to support that this relationship is statistically significant since the p-value is 0.788, largely exceeding 0.05 at 95% confidence level.
For (race)3, other, the estimated slope is -0.126, meaning that holding the other two factors constant, compared to being white, being in "other" racial group will result in a decrease of 0.126 units in satsoc. This relationship is also not statistically significant since it has a 0.051 p-value. It is closed to 0.05, but it is still insufficient to conclude any significance based on this value.

Factor 3: The slope between R's mental health&ability to think and satsoc is 0.540. Therefore, holding the other two factors constant, every unit of increase in one's score in hlthmntl will lead to increase of 0.540 in one's score in satsoc. Same as factor 1, both factors are coded inversely so lower hlthphys means better physical health. Based on the estimated slope, lower hlthphys will lead to lower satsoc, meaning that better mental health causes more social satisfaction. Since the p-value is 2e-16 at a high significance level, we can tell that this relationship is statistically significant.

R-squared: the adjusted R-squared is 0.37, meaning that 37% of the variance in a respondent's satsoc could be explained by the variance in their hlthphys, race, and mntlphys.

(g)Discussion

Although the relationships of sex vs. satsoc and race vs. satsoc are not proven significant in this project, it is undeniable that discrimination and inequalities based on sex or race still largely exist in our society. Therefore, the "insignificant" results here might be explicable: could some other social factors balance out the hypothetical negative effect that one's discriminated sex/race could have on their social satisfaction? Besides, from the model constructed, it is noteworthy that what affects one's social activities is not restricted to mental factor as commonly believed: one's physical status is also significantly important to one's social satisfaction. Following this, it will be sociologically meaningful to study the relationship between disabilities and one's engagement in social activities, and how exactly does the former influence the latter.