DS5110 Homework 4

Parth Shah

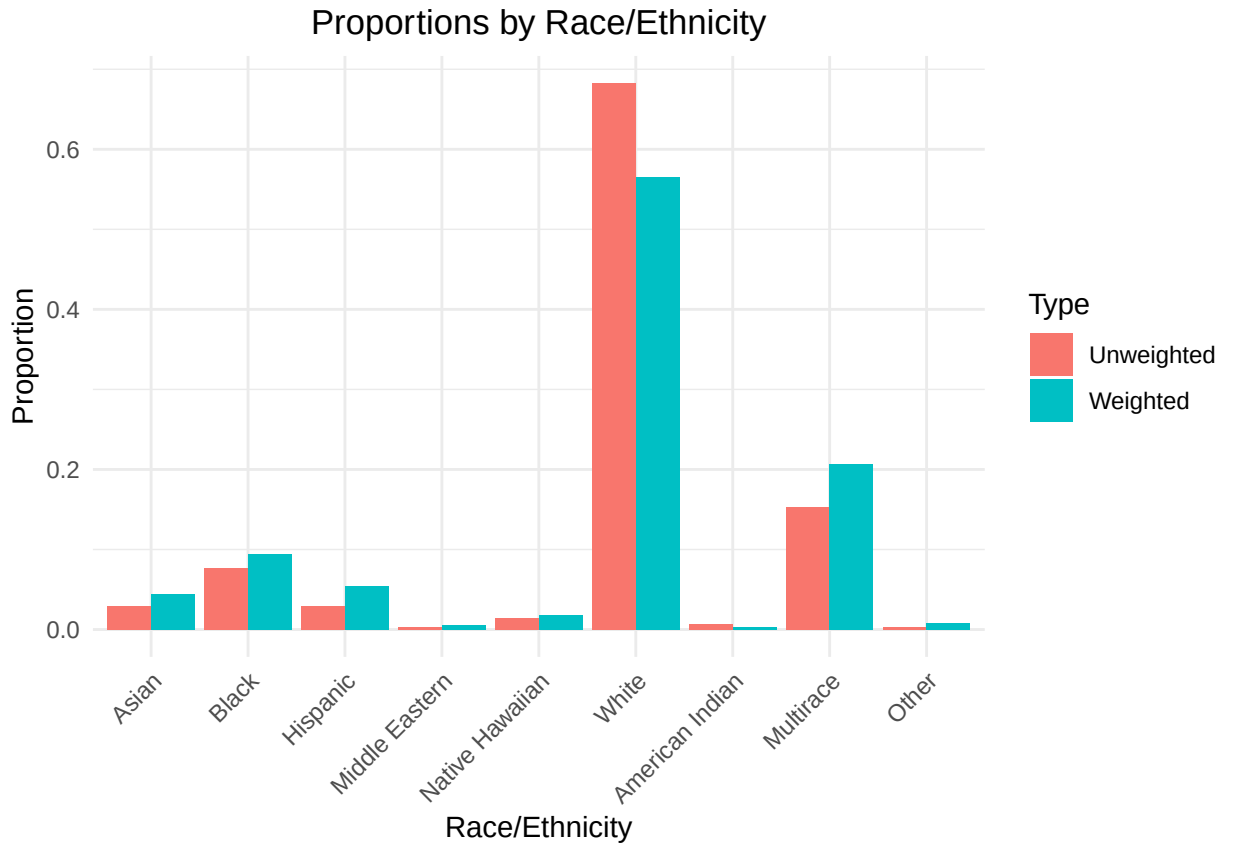
2023-03-19

Part A

Problem 1

```
suppressPackageStartupMessages(library(dplyr))
library(ggplot2)

load("37938-0001-Data.rda")
my_data <- da37938.0001

my_data <- my_data %>%
  mutate(RACE = recode(RACE,
                       "(1) Asian" = "Asian",
                       "(2) Black/AA" = "Black",
                       "(3) Hispanic/Latino" = "Hispanic",
                       "(4) Middle Eastern" = "Middle Eastern",
                       "(5) Native Hawaiian/Pacific Islander" = "Native Hawaiian",
                       "(6) White" = "White",
                       "(7) American Indian" = "American Indian",
                       "(8) Multirace" = "Multirace",
                       "(9) Other" = "Other"))

unweighted_props <- my_data %>%
  group_by(RACE) %>%
  summarize(count = n()) %>%
  mutate(Type = "Unweighted", prop = count / sum(count))

weighted_props <- my_data %>%
  group_by(RACE) %>%
  summarize(count = sum(WEIGHT))%>%
  mutate(Type = "Weighted", prop = count / sum(my_data$WEIGHT))

combined_props <- bind_rows(unweighted_props, weighted_props)

ggplot(combined_props, aes(x = RACE, y = prop, fill = Type)) +
  geom_col(position = position_dodge()) +
  ggtitle("Proportions by Race/Ethnicity") +
  xlab("Race/Ethnicity") +
  ylab("Proportion") +
  theme_minimal() +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```

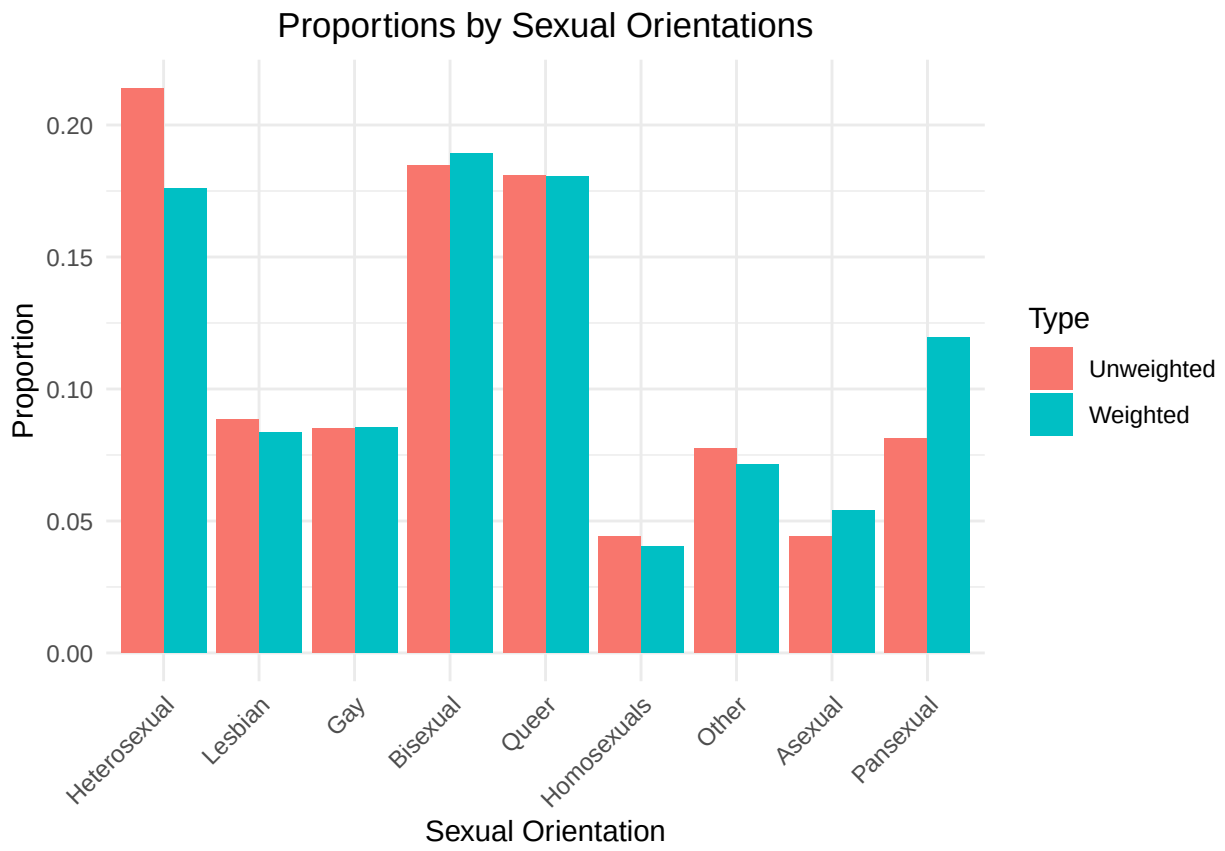


In the survey sample, several racial and ethnic groups are under-represented when compared to the population, including Asian, Black, Hispanic, Middle Eastern, Native Hawaiian, Multirace, and Other categories. On the other hand, the White and American Indian categories are over-represented in the survey sample. Overall, this indicates that the survey sample does not perfectly mirror the population's racial and ethnic composition, with some groups being over-represented while others are under-represented.

Problem 2

```
suppressPackageStartupMessages(library(dplyr))
library(ggplot2)

load("37938-0001-Data.rda")
my_data <- da37938.0001

my_data <- my_data %>%
  filter(!is.na(SEXUALID))

my_data <- my_data %>%
  mutate(SEXUALID = recode(SEXUALID,
                           "(1) Straight/heterosexual" = "Heterosexual",
                           "(2) Lesbian" = "Lesbian",
```

```
                        "(3) Gay" = "Gay",
                        "(4) Bisexual" = "Bisexual",
                        "(5) Queer" = "Queer",
                        "(6) Same-gender loving" = "Homosexuals",
                        "(7) Other" = "Other",
                        "(8) Asexual spectrum" = "Asexual",
                        "(9) Pansexual" = "Pansexual"))

unweighted_props <- my_data %>%
  group_by(SEXUALID) %>%
  summarize(count = n()) %>%
  mutate(Type = "Unweighted", prop = count / sum(count))

weighted_props <- my_data %>%
  group_by(SEXUALID) %>%
  summarize(count = sum(WEIGHT))%>%
  mutate(Type = "Weighted", prop = count / sum(my_data$WEIGHT))

combined_props <- bind_rows(unweighted_props, weighted_props)

ggplot(combined_props, aes(x = SEXUALID, y = prop, fill = Type)) +
  geom_col(position = position_dodge()) +
  ggtitle("Proportions by Sexual Orientations") +
  xlab("Sexual Orientation") +
  ylab("Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



In the survey sample, several sexual identities are over- or under-represented when compared to the population. Heterosexual, Lesbian, Queer, Homosexuals, and Other categories are over-represented in the survey sample compared to the population, as their unweighted proportions are higher than their respective weighted proportions. On the other hand, the Gay, Bisexual, Asexual, and Pansexual categories are under-represented in the survey sample, with their unweighted proportions being lower than their respective weighted proportions. Overall, the survey sample does not perfectly represent the population's sexual identity composition, with some sexual identities being over-represented and others being under-represented.

Part B

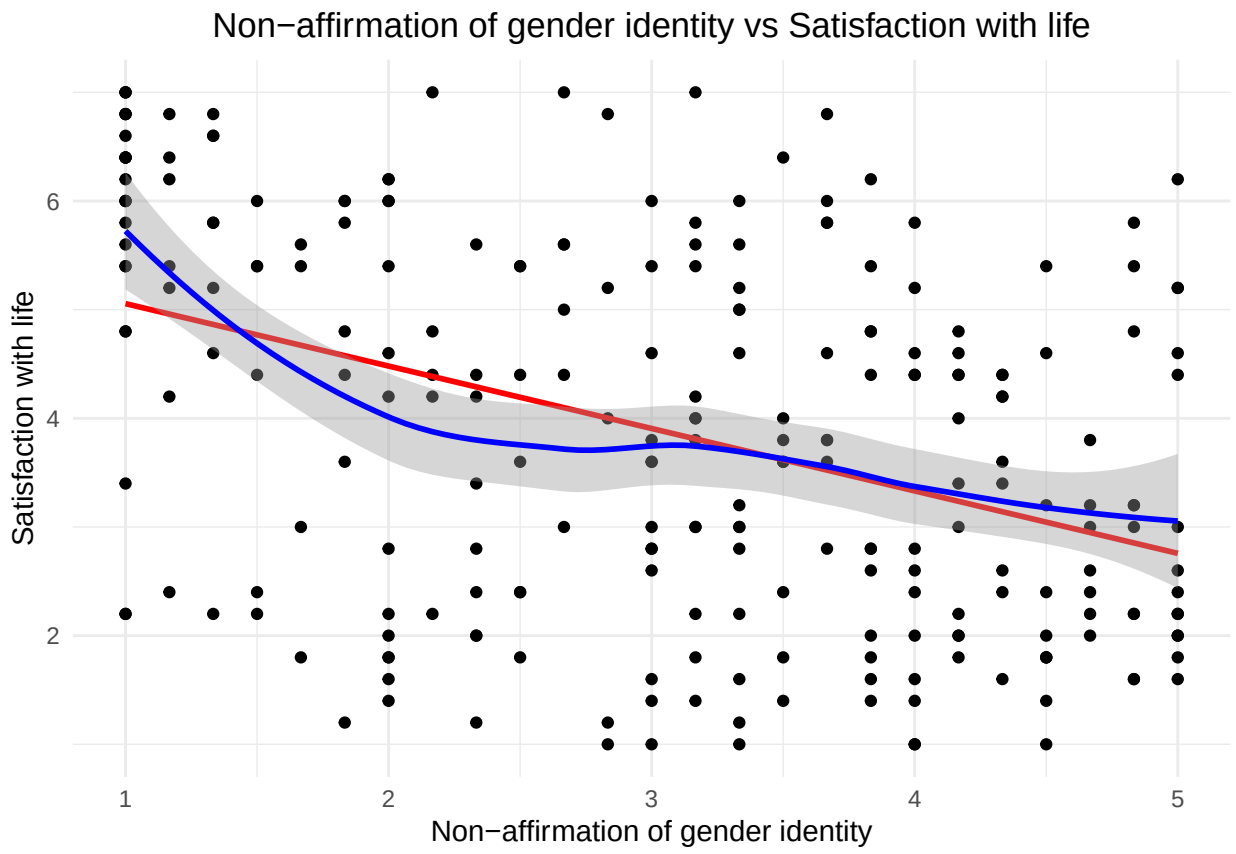
Problem 3

```
suppressPackageStartupMessages(library(dplyr))
library(ggplot2)

load("37938-0001-Data.rda")
my_data <- da37938.0001

my_data <- my_data %>% select(STUDYID, LIFESAT, LIFESAT_I, SOCIALWB, SOCIALWB_I,
                              NONAFFIRM, NONAFFIRM_I, NONDISCLOSURE,
                              NONDISCLOSURE_I, HCTHREAT, HCTHREAT_I,
                              KESSLER6, KESSLER6_I, EVERYDAY, EVERYDAY_I)
```

```r
my_data <- na.omit(my_data)

ggplot(my_data, aes(x = SOCIALWB_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Social well-being", y = "Satisfaction with life") +
  ggtitle("Satisfaction with life vs Social well-being") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

The plot shows a positive relationship between social well-being and life satisfaction as the slope is positive, indicating that an increase in social well-being leads to an increase in life satisfaction.

```r
ggplot(my_data, aes(x = NONAFFIRM_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Non-affirmation of gender identity", y = "Satisfaction with life") +
  ggtitle("Non-affirmation of gender identity vs Satisfaction with life") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

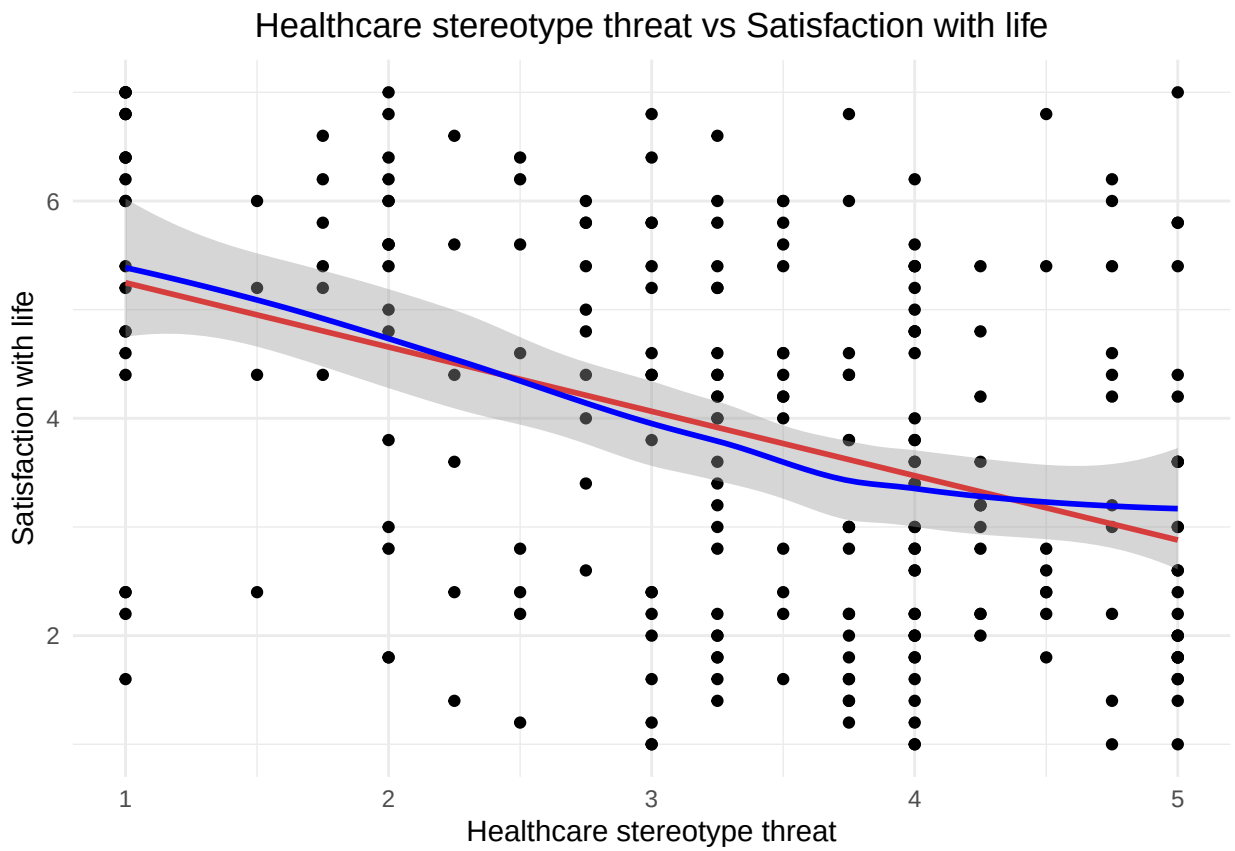


The plot shows a negative relationship between non-affirmation of gender identity and life satisfaction as the slope is negative, indicating that an increase in non-affirmation leads to a decrease in life satisfaction.

```
ggplot(my_data, aes(x = NONDISCLOSURE_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Non-disclosure of gender identity", y = "Satisfaction with life") +
  ggtitle("Non-disclosure of gender identity vs Satisfaction with life") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

The plot shows a negative relationship between non-disclosure of gender identity and life satisfaction as the slope is negative, indicating that an increase in non-disclosure leads to a decrease in life satisfaction.

```
ggplot(my_data, aes(x = HCTHREAT_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Healthcare stereotype threat", y = "Satisfaction with life") +
  ggtitle("Healthcare stereotype threat vs Satisfaction with life") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

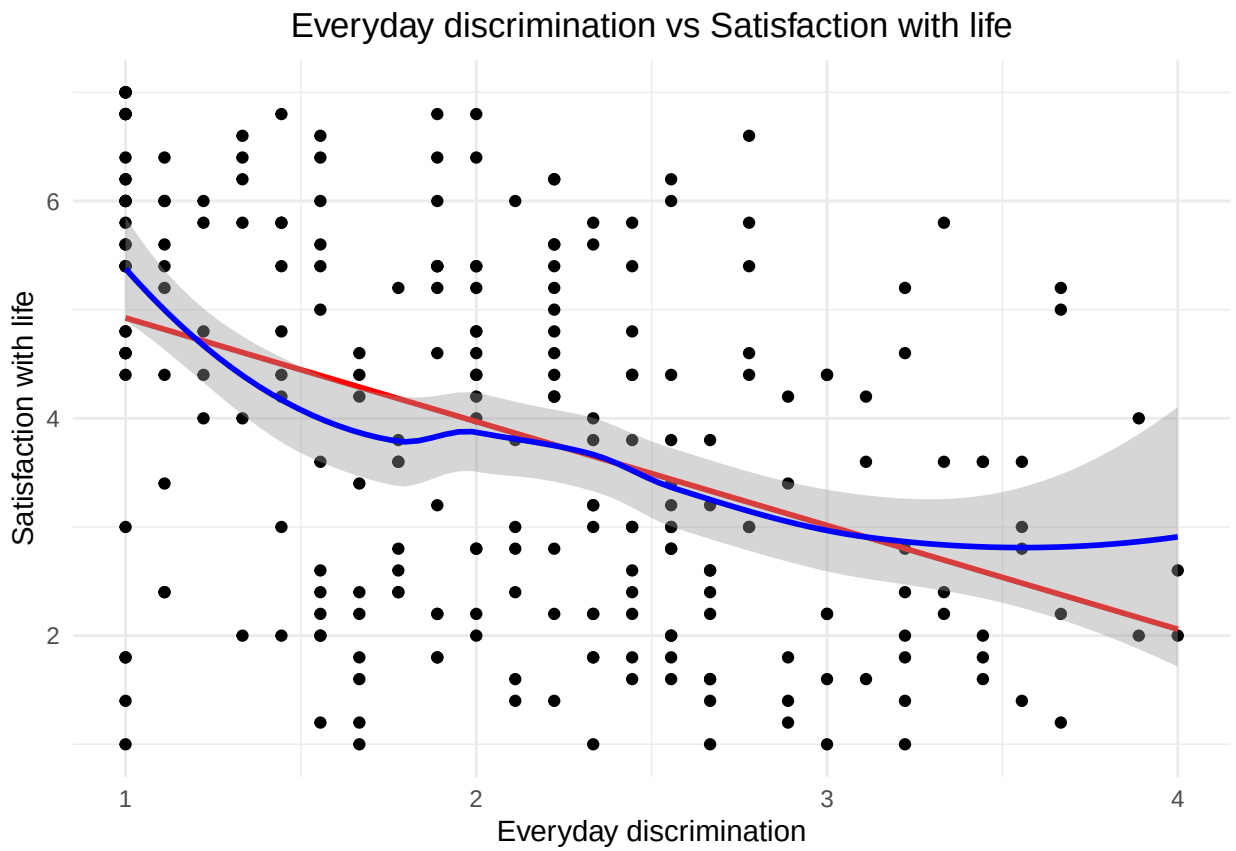


The plot shows a negative relationship between healthcare stereotype threat and life satisfaction as the slope is negative, indicating that an increase in healthcare stereotype threat leads to a decrease in life satisfaction.

```
ggplot(my_data, aes(x = KESSLER6_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Mental distress/disorder", y = "Satisfaction with life") +
  ggtitle("Mental distress/disorder vs Satisfaction with life") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

The plot shows a negative relationship between mental distress/disorder and life satisfaction as the slope is negative, indicating that an increase in mental distress/disorder leads to a decrease in life satisfaction.

```
ggplot(my_data, aes(x = EVERYDAY_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +
  labs(x = "Everyday discrimination", y = "Satisfaction with life") +
  ggtitle("Everyday discrimination vs Satisfaction with life") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



The plot shows a negative relationship between everyday discrimination and life satisfaction as the slope is negative, indicating that an increase in everyday discrimination leads to a decrease in life satisfaction.

Problem 4

From the above plots, mental distress/disorder exhibits a strong negative relationship with life satisfaction, as demonstrated by the steep negative slope in the plot. The slope indicates that as levels of mental distress or disorders increase, life satisfaction tends to decrease. This relationship suggests that mental distress/disorder is an important factor affecting life satisfaction and supports its inclusion as a predictor in a linear regression model. Additionally, the plot shows a relatively tight clustering of data points around the fitted line, indicating a strong correlation between the two variables, further justifying the inclusion of mental distress/disorder as a key predictor in the model.

Fit the model

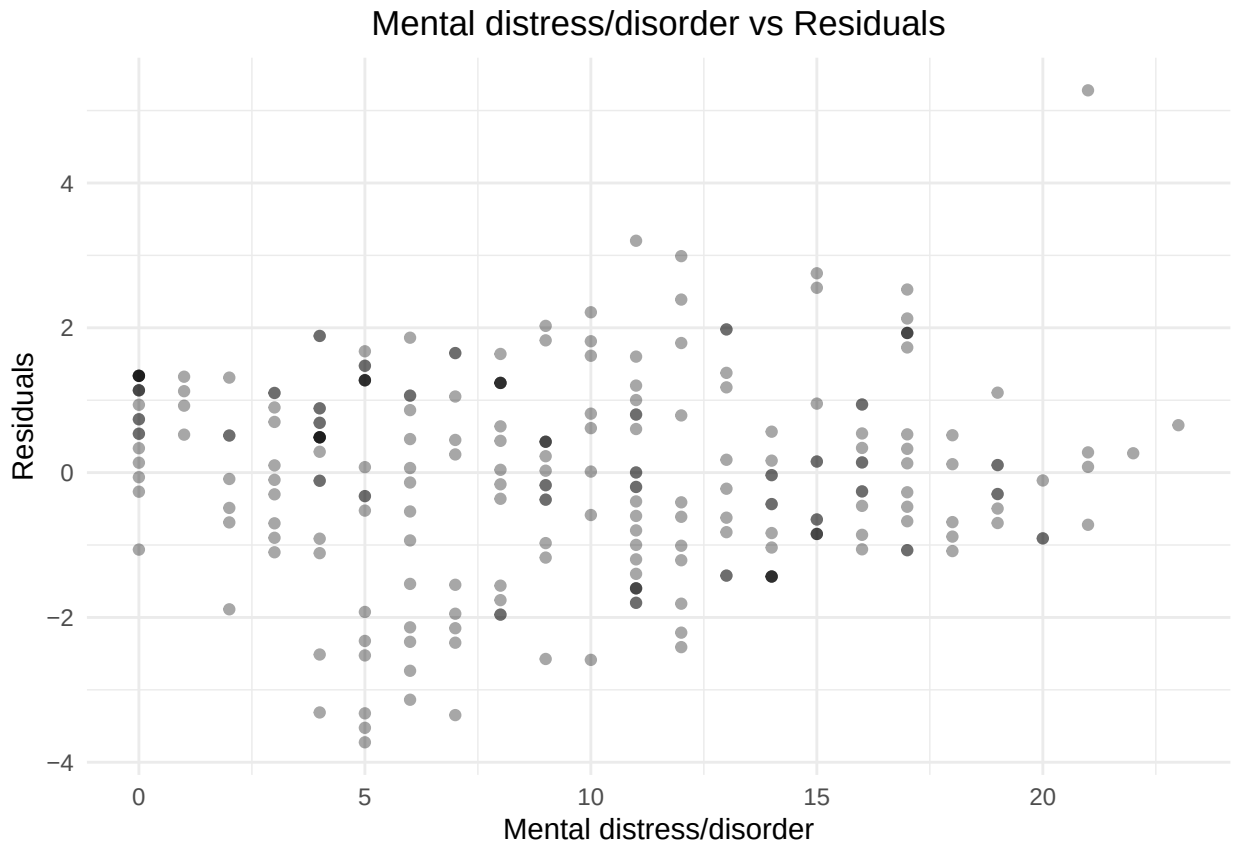
```
model1 <- lm(LIFESAT_I ~ KESSLER6_I, data = my_data)
summary(model1)
```

```
##
## Call:
## lm(formula = LIFESAT_I ~ KESSLER6_I, data = my_data)
##
## Residuals:
```

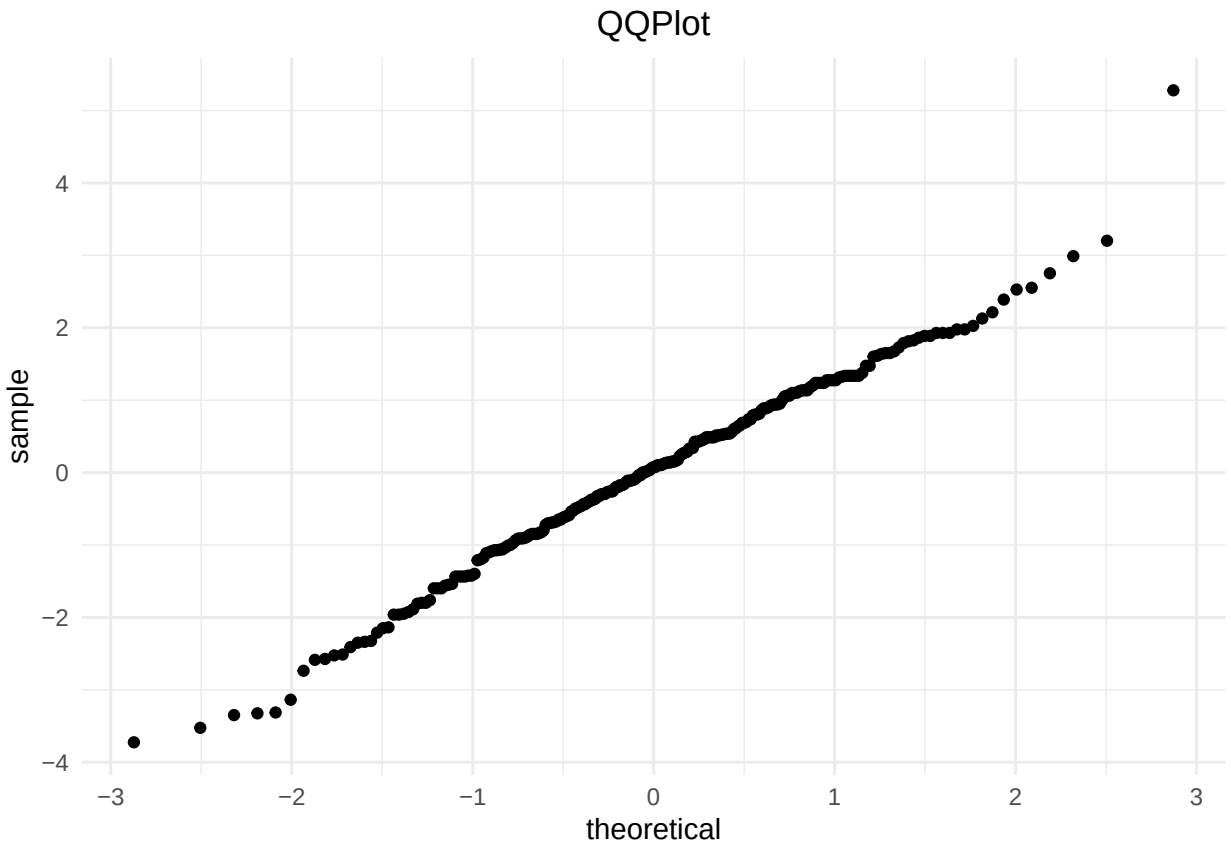
```
##     Min      1Q  Median      3Q     Max
## -3.7246 -0.8471  0.0754  0.9407  5.2794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.66336    0.16718   33.88   <2e-16 ***
## KESSLER6_I  -0.18775    0.01506  -12.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.357 on 243 degrees of freedom
## Multiple R-squared:   0.39,  Adjusted R-squared:  0.3875
## F-statistic: 155.4 on 1 and 243 DF,  p-value: < 2.2e-16
```

Model Diagnostics

```
library(modelr)

my_data %>%
  add_residuals(model1, "resid") %>%
  ggplot(aes(x = KESSLER6_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  labs(x = "Mental distress/disorder", y = "Residuals") +
  ggtitle("Mental distress/disorder vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
my_data %>%
  add_residuals(model1, "resid") %>%
  ggplot(aes(sample=resid)) +
  geom_qq() +
  ggtitle("QQPlot") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



There is a data point with a residual greater than 4, which is significantly higher than the residuals for the other data points, this could be an outlier. Outliers can have a substantial impact on the model, potentially influencing the slope and intercept of the fitted line and leading to less accurate predictions. Hence, I have removed that point from the data and re-fitted the model, and performed the model diagnostics again.

Re-fit the model after removing outlier

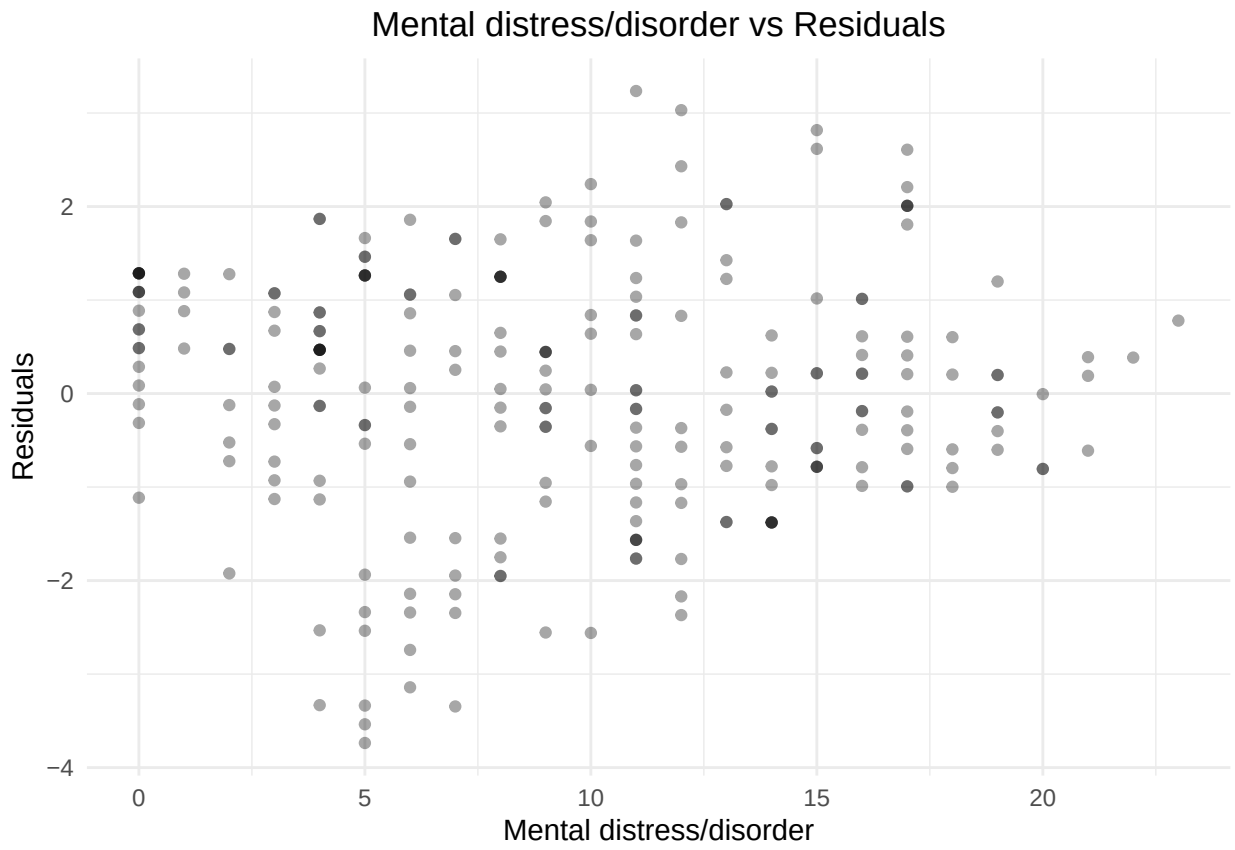
```
my_data <- my_data %>%
  add_residuals(model1, "resid") %>%
  filter(resid <= 4)
new_model1 <- lm(LIFESAT_I ~ KESSLER6_I, data = my_data)
summary(new_model1)
```

```
##
## Call:
## lm(formula = LIFESAT_I ~ KESSLER6_I, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7369 -0.7841  0.0608  0.9177  3.2354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.71386    0.16259   35.14   <2e-16 ***
```

```
## KESSLER6_I  -0.19539    0.01473  -13.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.316 on 242 degrees of freedom
## Multiple R-squared:  0.4211, Adjusted R-squared:  0.4187
## F-statistic:   176 on 1 and 242 DF,  p-value: < 2.2e-16
```

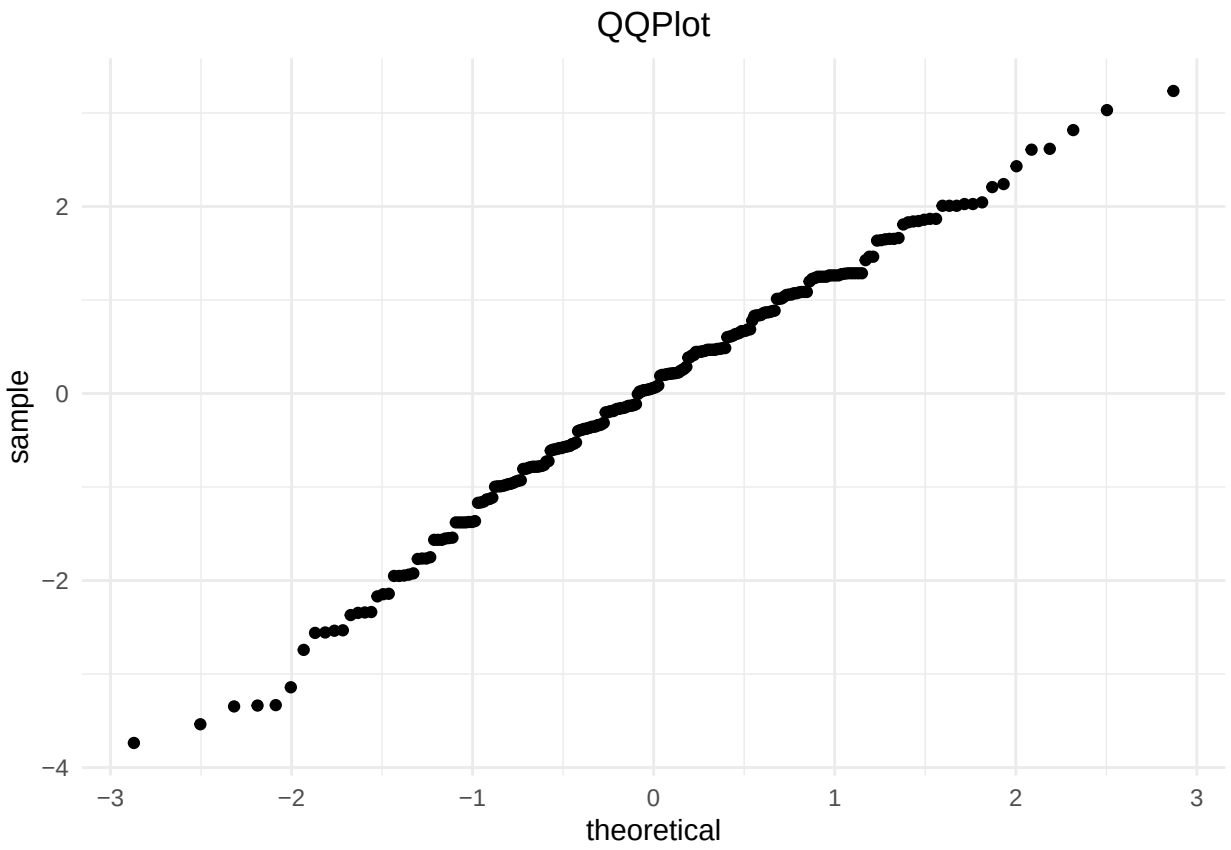
Model Diagnostics

```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = KESSLER6_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  labs(x = "Mental distress/disorder", y = "Residuals") +
  ggtitle("Mental distress/disorder vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



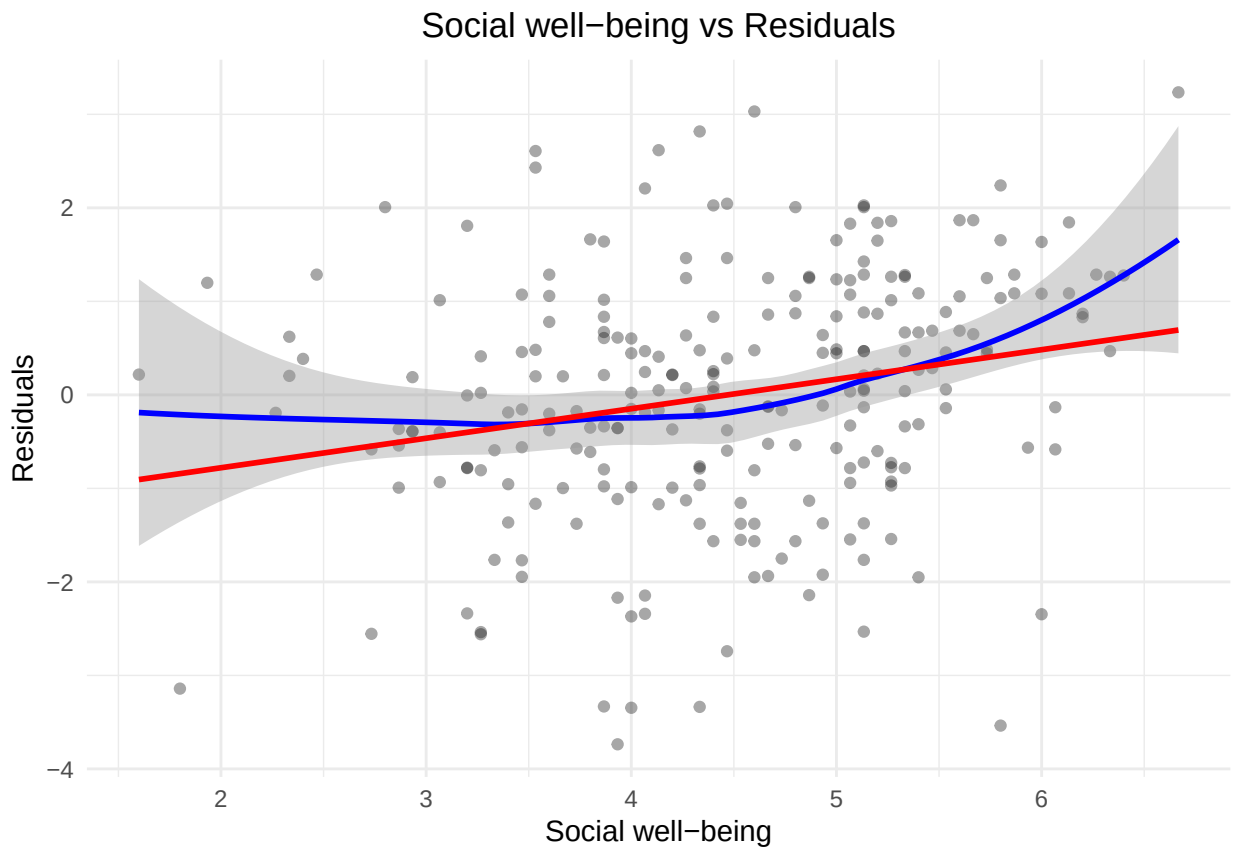

```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(sample=resid)) +
  geom_qq() +
  ggtitle("QQPlot") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

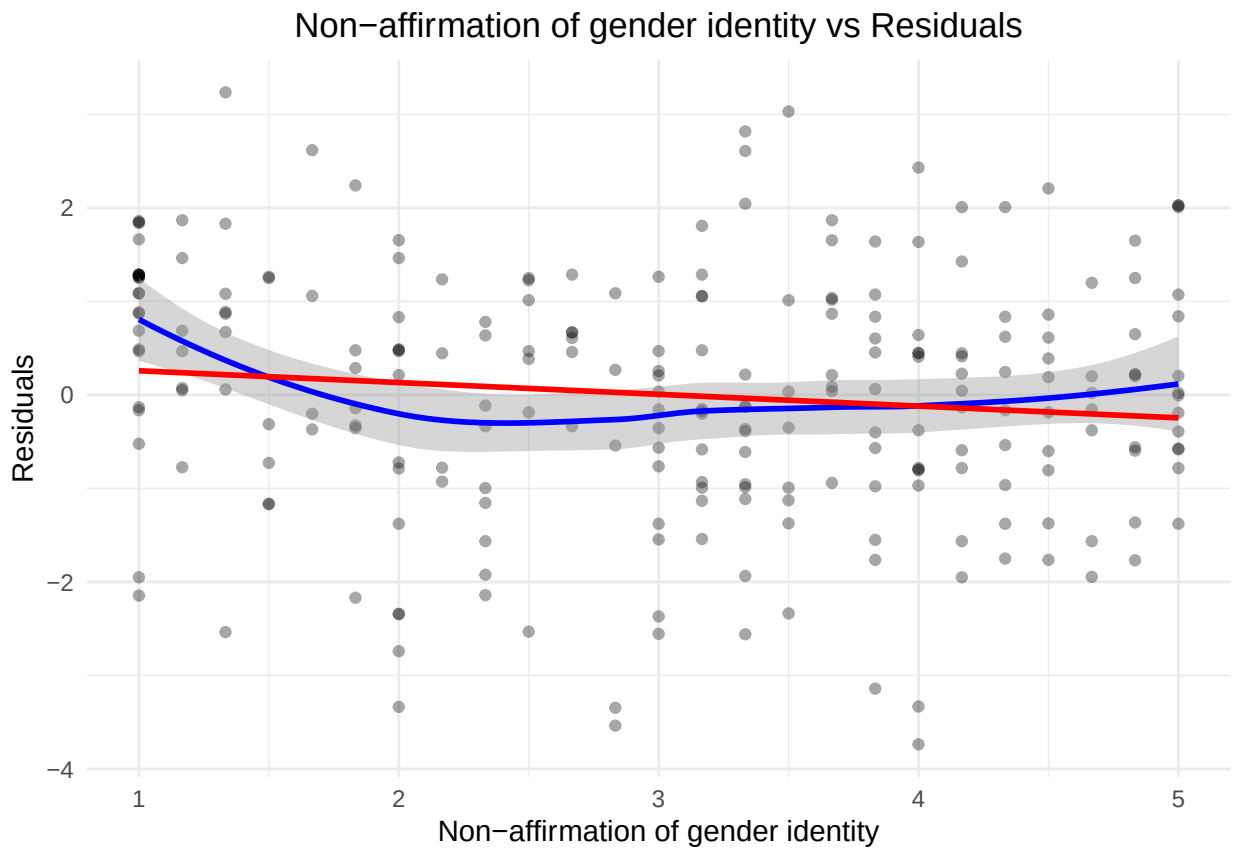


Problem 5

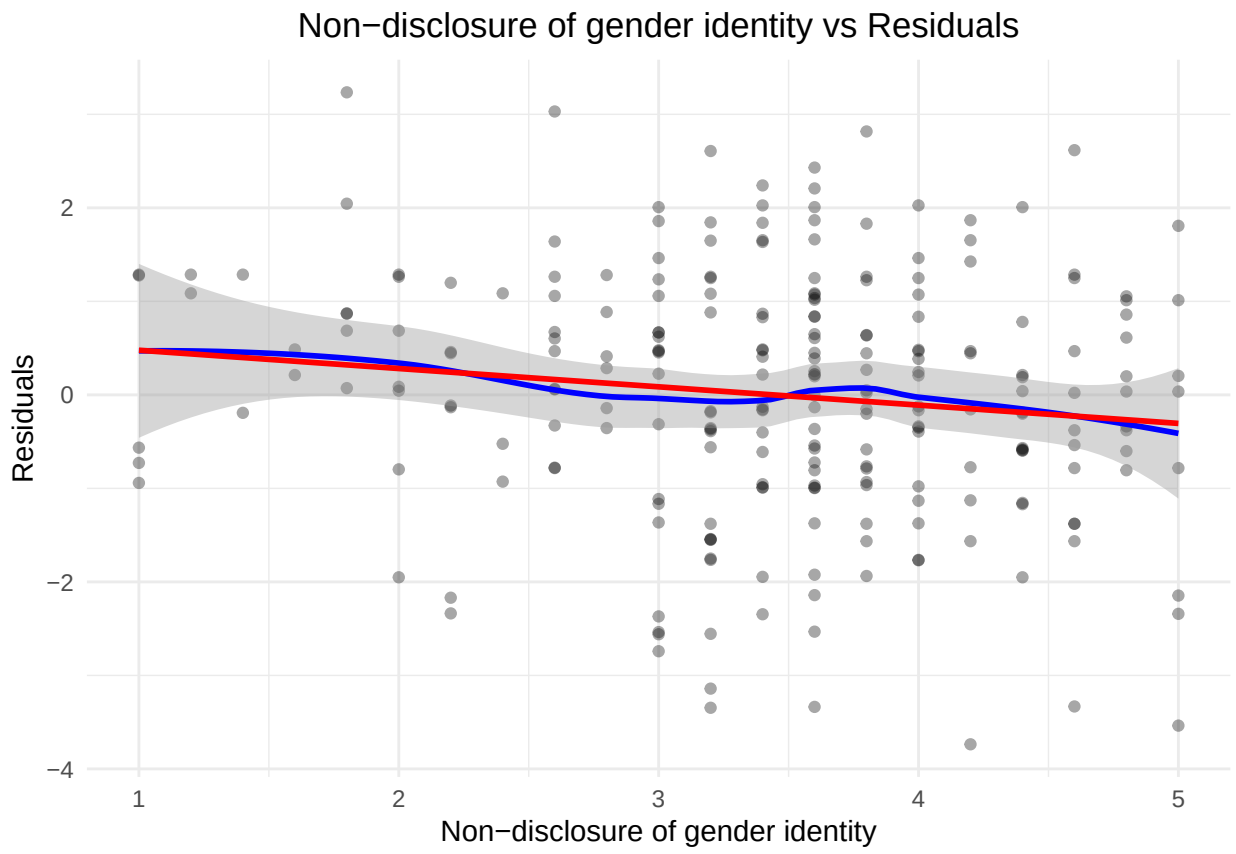
```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = SOCIALWB_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +
  labs(x = "Social well-being", y = "Residuals") +
  ggtitle("Social well-being vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



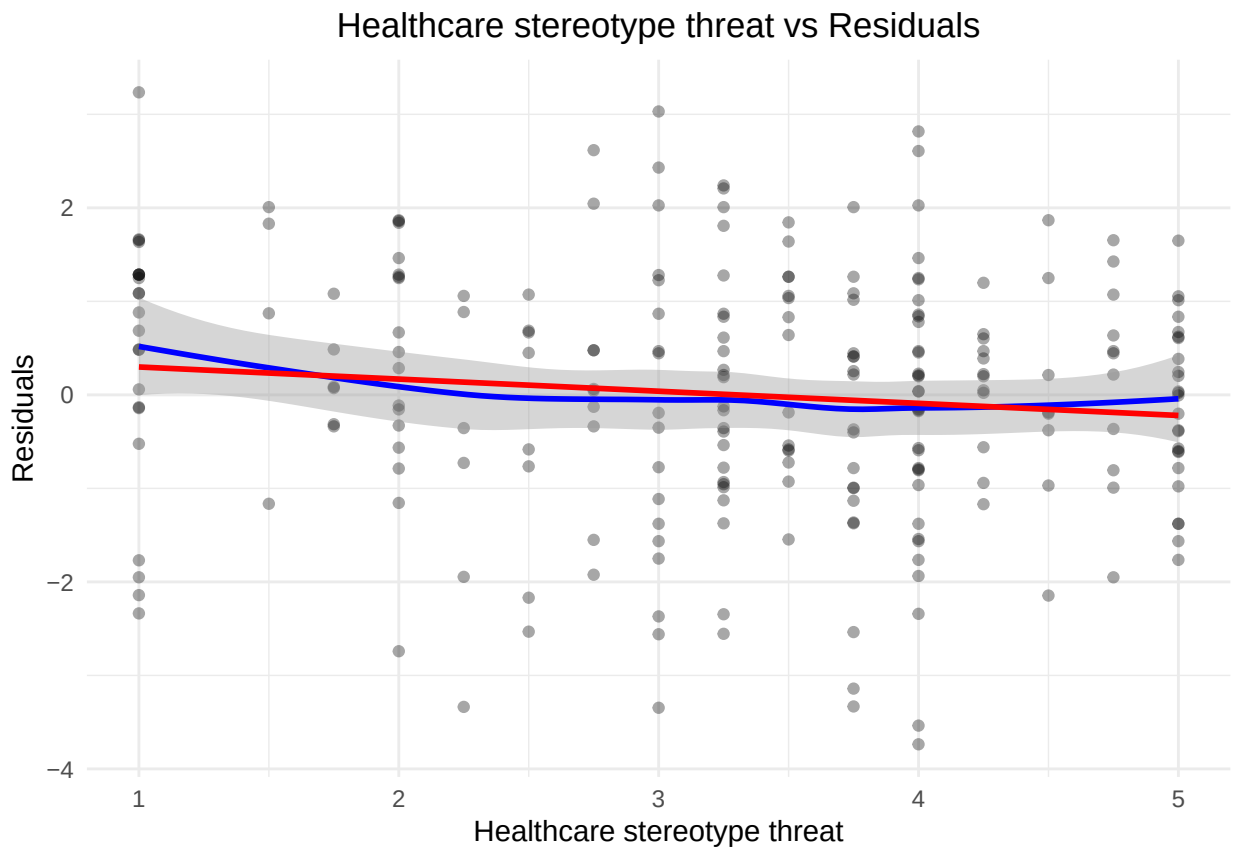

```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = NONAFFIRM_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +
  labs(x = "Non-affirmation of gender identity", y = "Residuals") +
  ggtitle("Non-affirmation of gender identity vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



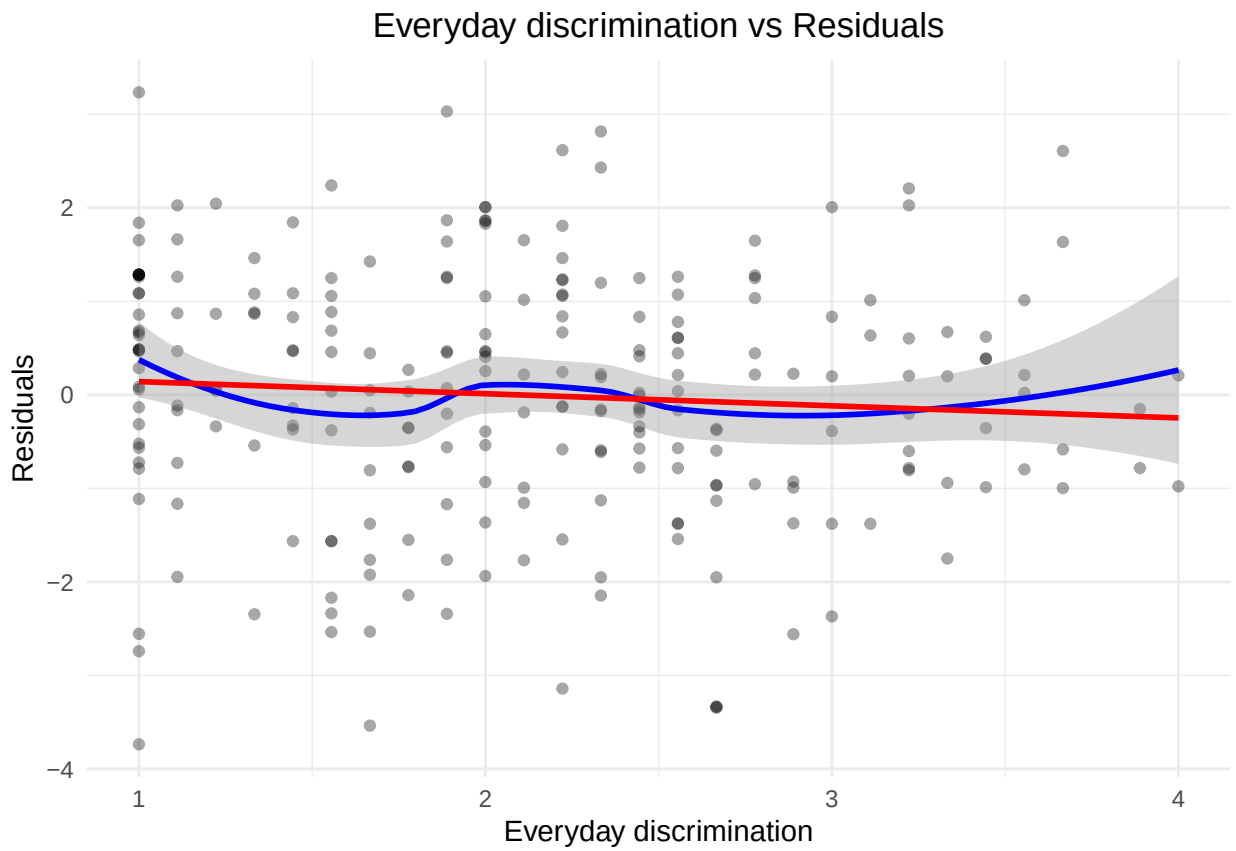

```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = NONDISCLOSURE_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +
  labs(x = "Non-disclosure of gender identity", y = "Residuals") +
  ggtitle("Non-disclosure of gender identity vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = HCTHREAT_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +
  labs(x = "Healthcare stereotype threat", y = "Residuals") +
  ggtitle("Healthcare stereotype threat vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
my_data %>%
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = EVERYDAY_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +
  labs(x = "Everyday discrimination", y = "Residuals") +
  ggtitle("Everyday discrimination vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the above plots, I decided to use Social well-being (SOCIALWB_I) alongside KESSLER6_I (Mental distress/disorder) in my model, as I noticed a higher positive slope and a trend line close to the linear line. This indicates that SOCIALWB_I has a strong positive relationship with life satisfaction (LIFESAT_I) and can provide valuable insights. By incorporating predictors with varying relationships to the outcome, like the positive link with SOCIALWB_I and the negative one with KESSLER6_I, I can better understand the complex factors influencing life satisfaction, taking into account both the negative effects of mental distress and the positive contributions of social well-being.

Fit the model

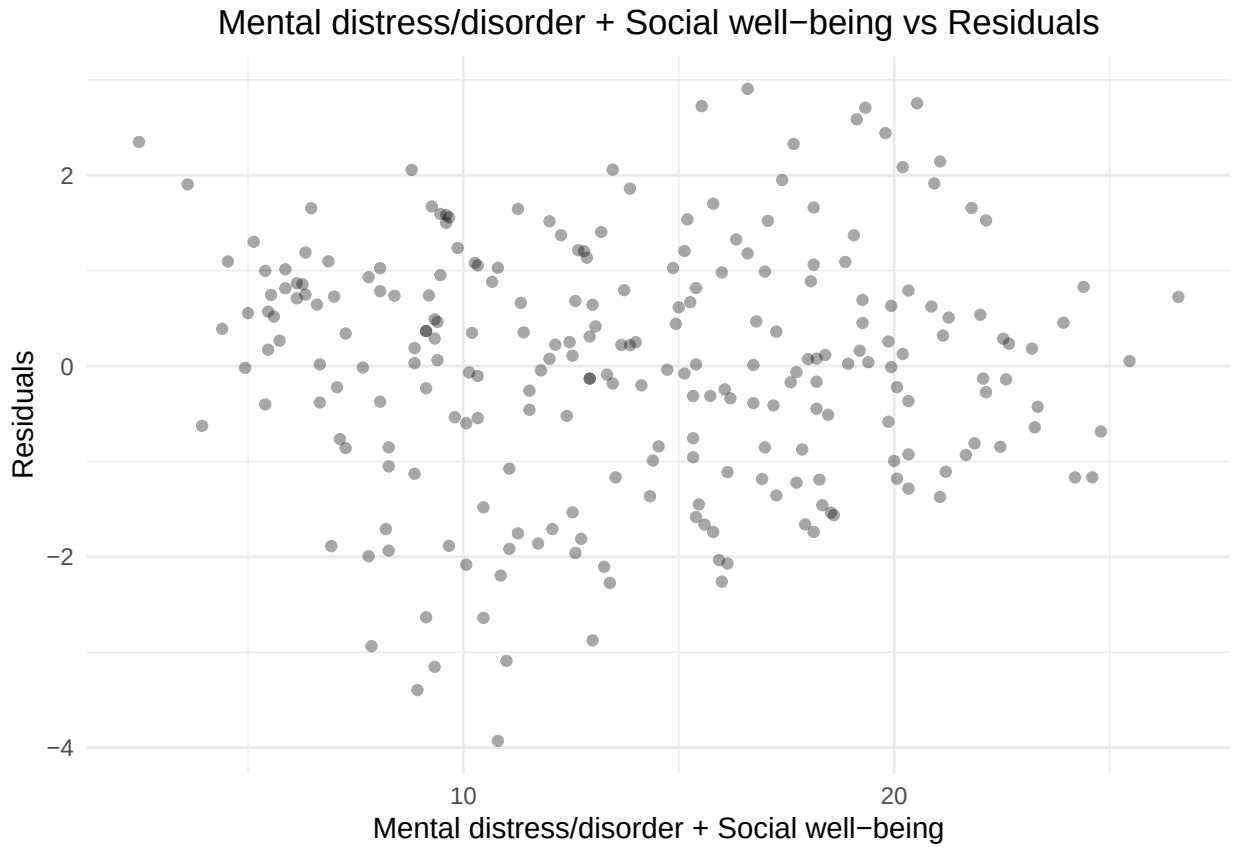
```
model2 <- lm(LIFESAT_I ~ KESSLER6_I + SOCIALWB_I, data = my_data)
summary(model2)
```

```
##
## Call:
## lm(formula = LIFESAT_I ~ KESSLER6_I + SOCIALWB_I, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9290 -0.8520  0.0763  0.8362  2.9047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.68373    0.51227   7.191 8.06e-12 ***
```

```
## KESSLER6_I  -0.16612    0.01589 -10.453  < 2e-16 ***
## SOCIALWB_I   0.39239    0.09422   4.164 4.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 241 degrees of freedom
## Multiple R-squared:   0.46,  Adjusted R-squared:  0.4555
## F-statistic: 102.6 on 2 and 241 DF,  p-value: < 2.2e-16
```

Model Diagnostics

```
my_data %>%
  add_residuals(model2, "resid") %>%
  ggplot(aes(x = KESSLER6_I + SOCIALWB_I)) +
  geom_point(aes(y = resid), alpha = 0.35) +
  labs(x = "Mental distress/disorder + Social well-being", y = "Residuals") +
  ggtitle("Mental distress/disorder + Social well-being vs Residuals") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
my_data %>%
  add_residuals(model2, "resid") %>%
  ggplot(aes(sample=resid)) +
  geom_qq() +
```

```
  ggtitle("QQPlot") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```