DS5110 Homework 4

Ameya Santosh Gidh

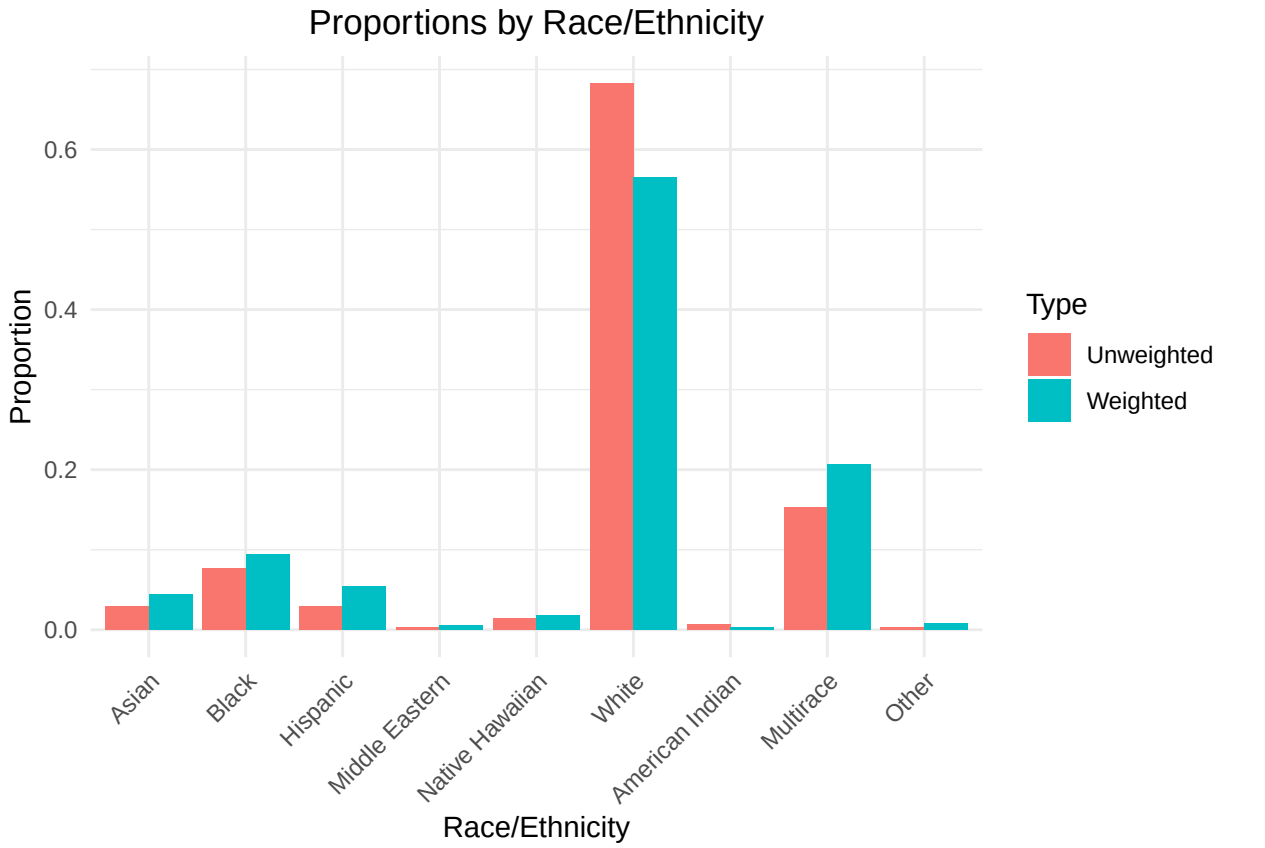
2024-03-09

Part A

Problem 1

```r
# Load necessary libraries
suppressPackageStartupMessages(library(dplyr))  # Suppress library startup messages
library(ggplot2)  # Load ggplot2 library

# Load the dataset
load("37938-0001-Data.rda")
my_data <- da37938.0001  # Assign dataset to my_data variable

# Recode race categories for clarity
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"))

# Calculate unweighted proportions by race
unweighted_props <- my_data %>%
  group_by(RACE) %>%
  summarize(count = n()) %>%
  mutate(Type = "Unweighted", prop = count / sum(count))

# Calculate weighted proportions by race
weighted_props <- my_data %>%
  group_by(RACE) %>%
  summarize(count = sum(WEIGHT)) %>%
  mutate(Type = "Weighted", prop = count / sum(my_data$WEIGHT))

# Combine unweighted and weighted proportions
combined_props <- bind_rows(unweighted_props, weighted_props)

# Create a bar plot showing proportions by race/ethnicity
ggplot(combined_props, aes(x = RACE, y = prop, fill = Type)) +
```

```r
  geom_col(position = position_dodge()) +  # Plot bars with dodge positioning
  ggtitle("Proportions by Race/Ethnicity") +  # Set plot title
  xlab("Race/Ethnicity") +  # Set x-axis label
  ylab("Proportion") +  # Set y-axis label
  theme_minimal() +  # Set minimal theme for the plot
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        plot.title = element_text(hjust = 0.5))  # Center plot title
```



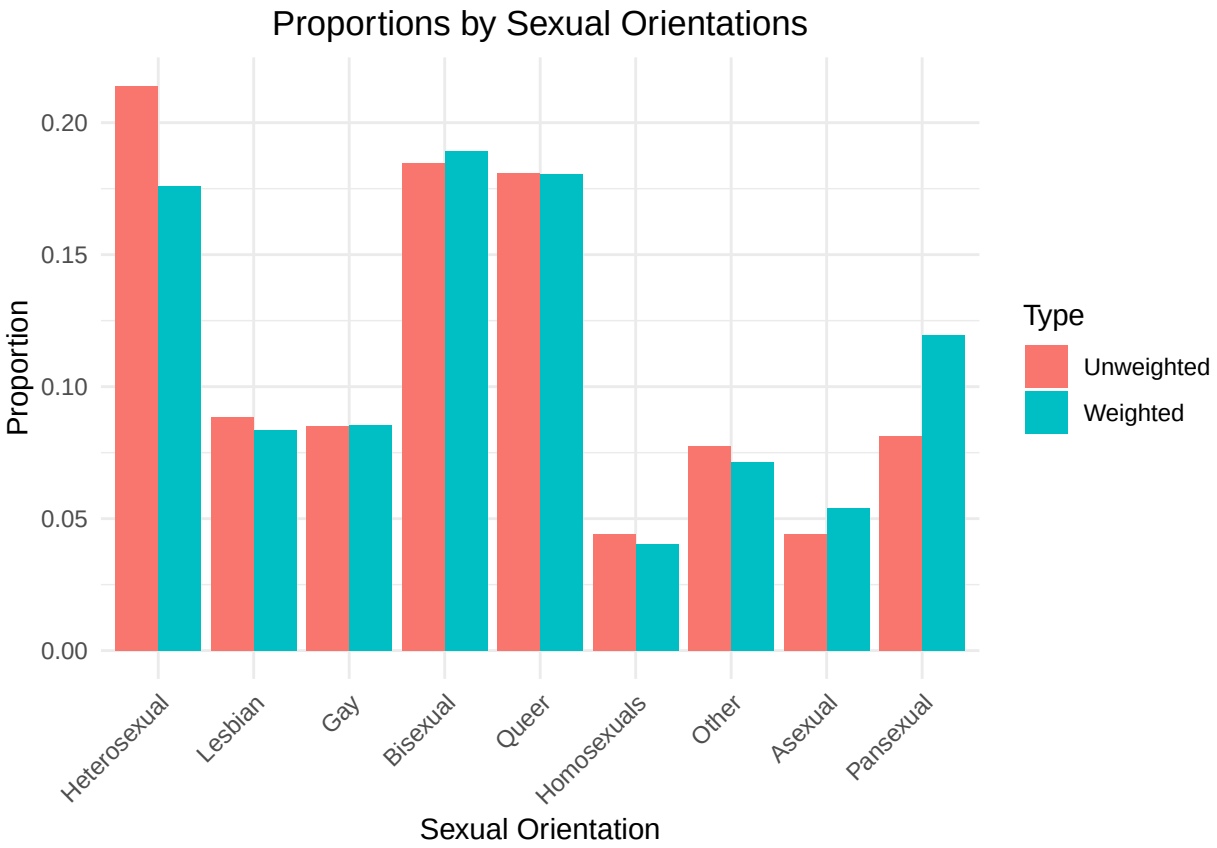
Comments and Observations: 1. The survey sample shows an over-representation of the White race and American Indian race compared to the population. This is evident as the unweighted proportion exceeds the weighted proportion. 2. Conversely, the Asian, Black/AA, Hispanic/Latino, Middle Eastern, Multirace, Native Hawaiian/Pacific Islander, and Other races are underrepresented in the survey sample relative to the population. This is indicated by the weighted proportion being higher than the unweighted proportion. 3. These observations suggest the presence of sampling bias and unit non-response. 4. The White race and Multirace exhibit significantly higher proportions in both weighted and unweighted proportions compared to other races in the population. 5. Conversely, the American Indian, Middle Eastern, and Other races display significantly lower proportions in both weighted and unweighted proportions compared to other races in the population.

The survey sample does not accurately reflect the racial and ethnic composition of the population. Certain groups are over-represented while others are under-represented, indicating discrepancies between the sample and the population.

Problem 2

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

```r
# Load the dataset
load("37938-0001-Data.rda")
my_data <- da37938.0001  # Assign dataset to my_data variable

# Filter out rows with missing values in the SEXUALID column
my_data <- my_data %>%
  filter(!is.na(SEXUALID))

# Recode sexual orientation categories for clarity
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"))

# Calculate unweighted proportions by sexual orientation
unweighted_props <- my_data %>%
  group_by(SEXUALID) %>%
  summarize(count = n()) %>%
  mutate(Type = "Unweighted", prop = count / sum(count))

# Calculate weighted proportions by sexual orientation
weighted_props <- my_data %>%
  group_by(SEXUALID) %>%
  summarize(count = sum(WEIGHT))%>%
  mutate(Type = "Weighted", prop = count / sum(my_data$WEIGHT))

# Combine unweighted and weighted proportions
combined_props <- bind_rows(unweighted_props, weighted_props)

# Create a bar plot showing proportions by sexual orientation
ggplot(combined_props, aes(x = SEXUALID, y = prop, fill = Type)) +
  geom_col(position = position_dodge()) +  # Plot bars with dodge positioning
  ggtitle("Proportions by Sexual Orientations") +  # Set plot title
  xlab("Sexual Orientation") +  # Set x-axis label
  ylab("Proportion") +  # Set y-axis label
  theme_minimal() +  # Set minimal theme for the plot
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        plot.title = element_text(hjust = 0.5))  # Center plot title
```



In the survey sample, there are discrepancies in the representation of various sexual identities compared to the population. Heterosexual, Lesbian, Queer, Homosexuals, and Other categories are over-represented in the sample, as their unweighted proportions exceed their respective weighted proportions. Conversely, the Gay, Bisexual, Asexual, and Pansexual categories are under-represented, with their unweighted proportions being lower than their weighted counterparts. This indicates that the survey sample does not accurately reflect the sexual identity composition of the population, with some identities being over-represented while others are under-represented.

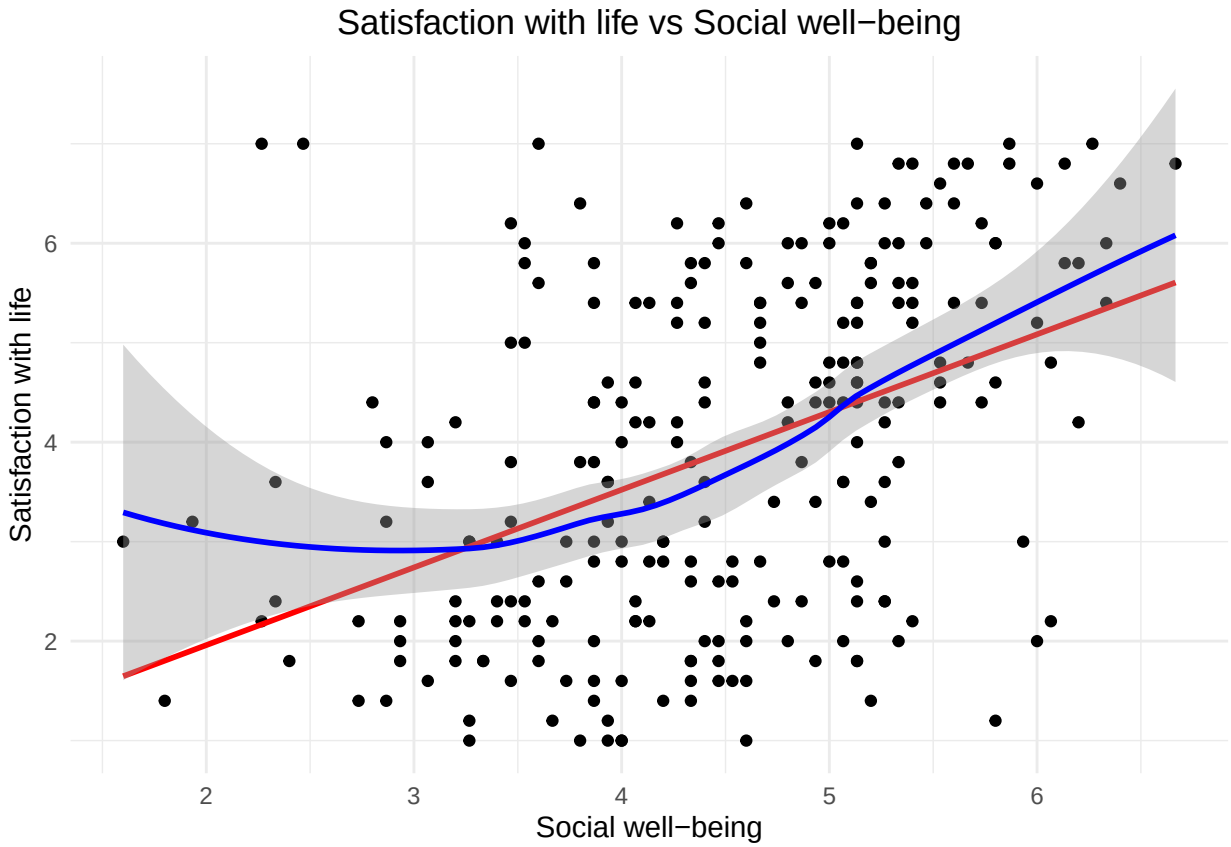
Comments and Observations: 1. The survey sample shows over-representation of the White and American Indian races compared to the population, as indicated by higher unweighted proportions than weighted proportions. 2. Conversely, the Asian, Black/AA, Hispanic/Latino, Middle Eastern, Multirace, Native Hawaiian/Pacific Islander, and Other races are under-represented in the survey sample compared to the population, with weighted proportions higher than unweighted proportions. 3. These observations imply the presence of sampling bias and unit non-response. 4. The White race and Multirace exhibit significantly higher proportions in both weighted and unweighted measures compared to other races in the population. 5. Conversely, the American Indian, Middle Eastern, and Other races display significantly lower proportions in both weighted and unweighted measures compared to other races in the population.

Part B

Problem 3

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

load("37938-0001-Data.rda")
# Assign dataset to my_data variable
```

```r
my_data <- da37938.0001

# Select relevant columns from the dataset
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)

# Remove rows with missing values
my_data <- na.omit(my_data)

# Create a scatter plot to visualize the relationship between satisfaction with life and social well-be
ggplot(my_data, aes(x = SOCIALWB_I, y = LIFESAT_I)) +  # Define x and y variables
  geom_point() +  # Plot points
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +  # Add linear regression lin
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ x') +  # Add smoothed line
  labs(x = "Social well-being", y = "Satisfaction with life") +  # Set x and y axis labels
  ggtitle("Satisfaction with life vs Social well-being") +  # Set plot title
  theme_minimal() +  # Set minimal theme for the plot
  theme(plot.title = element_text(hjust = 0.5))  # Center plot title
```

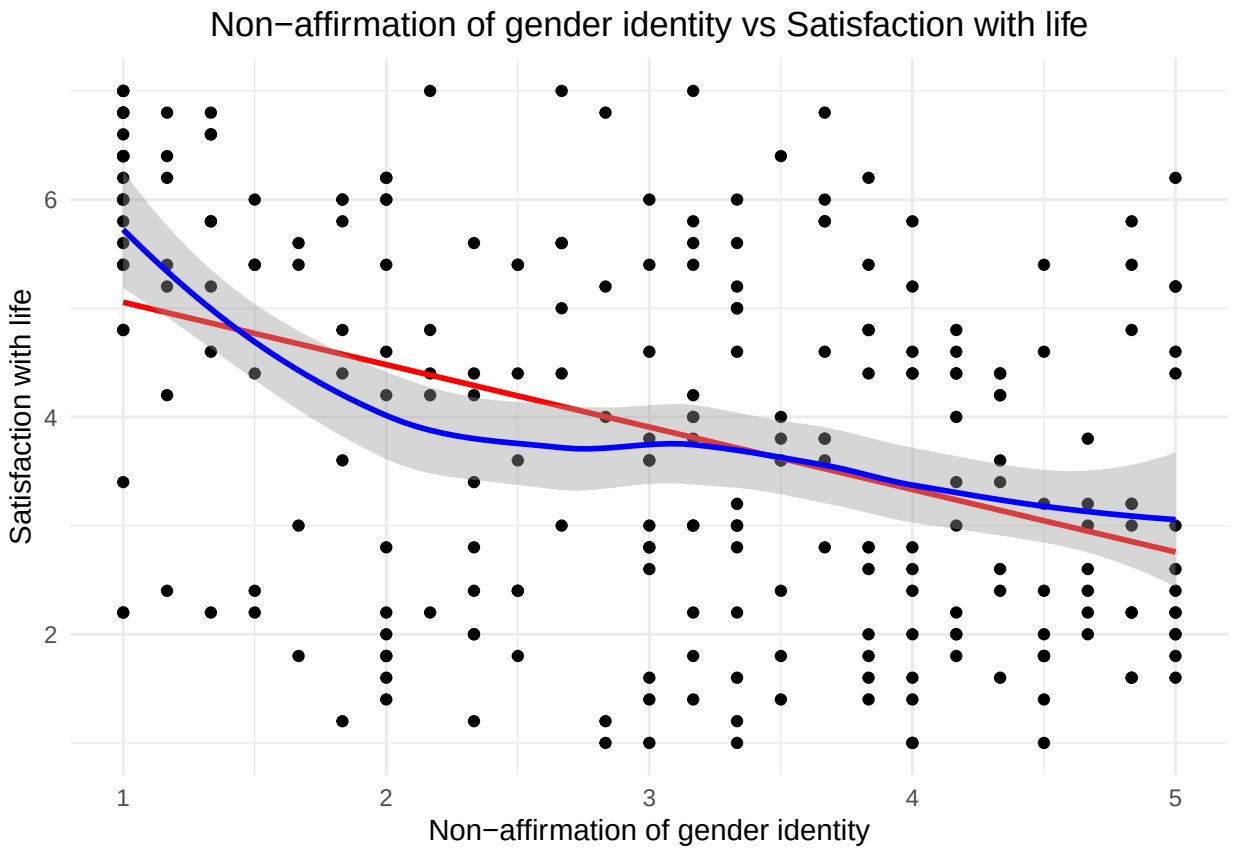


Visualization Comments:

1. The plot clearly illustrates a strong positive relationship between social well-being and life satisfaction.
2. The positive slope of the plot indicates that as social well-being increases, there is a corresponding increase in life satisfaction.
3. It demonstrates a directly positive and linearly increasing relationship between well-being and life

satisfaction.

```
# Plot relationship between non-affirmation of gender identity and life satisfaction
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))
```



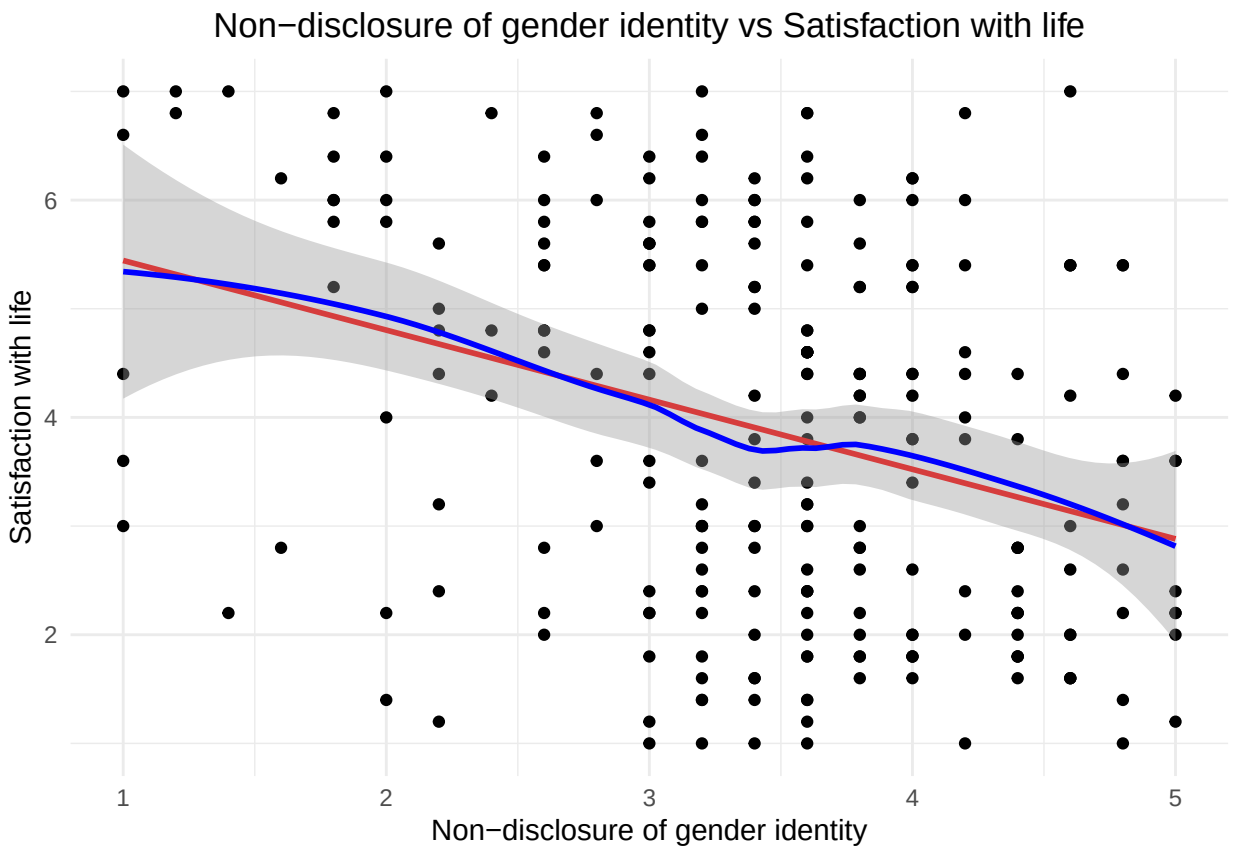
Comments:

1. The plot clearly demonstrates a negative relationship between non-affirmation of gender identity and life satisfaction.
2. The negative slope of the plot suggests that an increase in non-affirmation results in a decrease in life satisfaction.

Overall, the visualization effectively communicates that higher levels of non-affirmation of gender identity are associated with lower levels of life satisfaction.

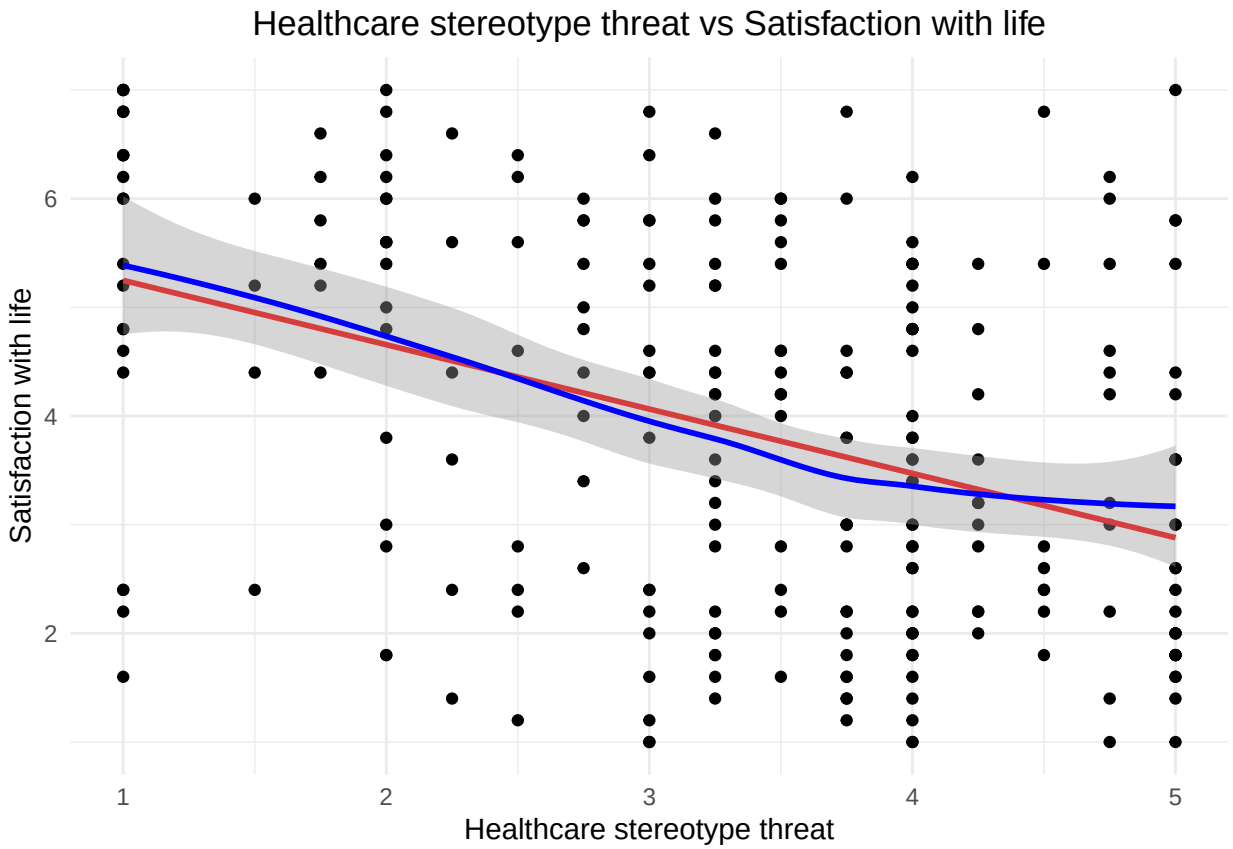
```
# Plot relationship between non-disclosure of gender identity and life satisfaction
ggplot(my_data, aes(x = NONDISCLOSURE_I, y = LIFESAT_I)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'lm', formula = 'y ~ x', color="red") +
  geom_smooth(method = 'loess', color = "blue", formula = 'y ~ 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))
```



Comments: 1. The plot clearly illustrates a negative relationship between non-disclosure of gender identity and life satisfaction. As the x-axis variable (non-disclosure of gender identity) increases, the y-axis variable (satisfaction with life) tends to decrease. This negative correlation is evident from the downward slope of the regression lines. 2. The negative slope of the linear regression line further emphasizes the inverse relationship between non-disclosure of gender identity and life satisfaction. It indicates that an increase in non-disclosure tends to be associated with a decrease in life satisfaction, supporting the notion that concealing one's gender identity may have adverse effects on overall satisfaction with life.

```
# Plot relationship between healthcare stereotype threat and 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))
```

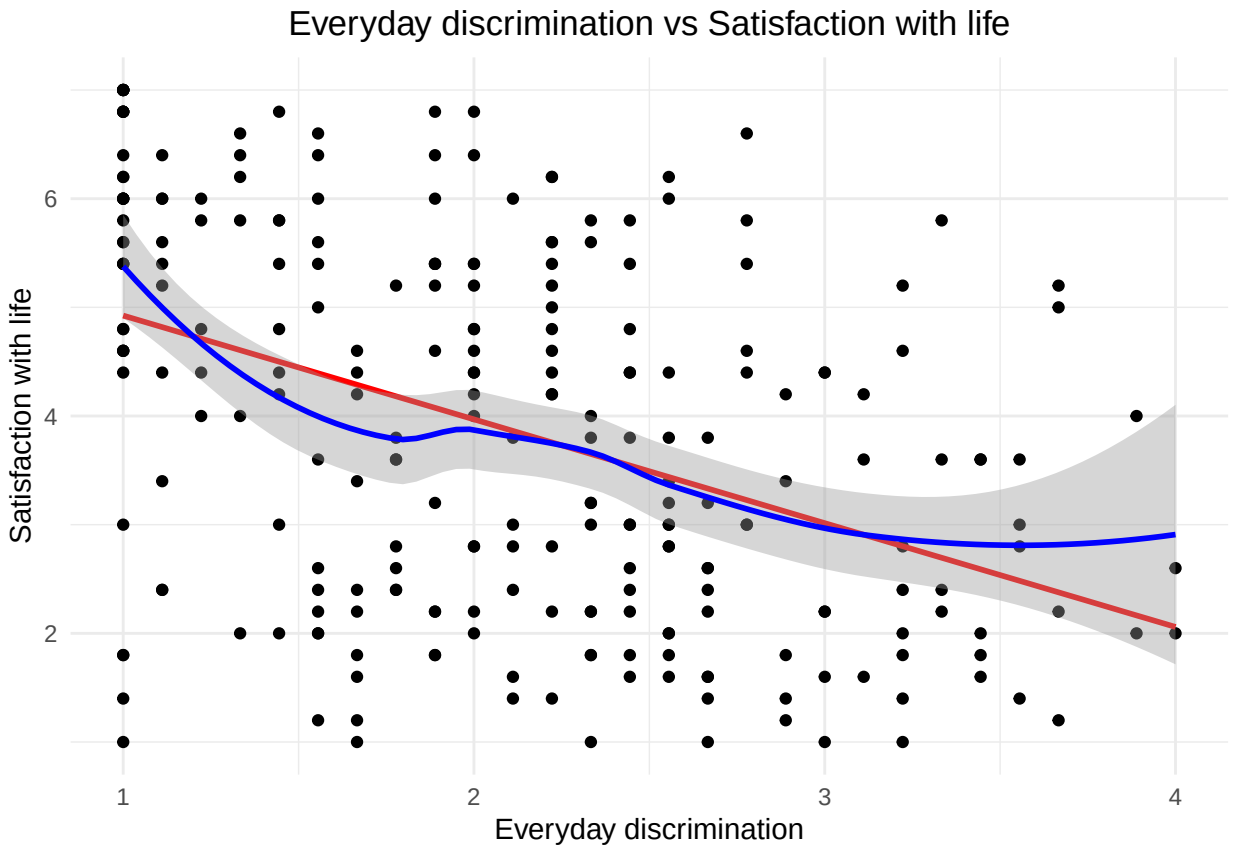


Comments: 1. The plot clearly illustrates a strong negative relationship between healthcare stereotype threat and life satisfaction. 2. The negative slope of the plot indicates that an increase in healthcare stereotype threat results in a decrease in life satisfaction. Overall, the visualization effectively conveys that higher levels of healthcare stereotype threat are associated with lower levels of life satisfaction.

```
# Plot relationship between mental distress/disorder and 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))
```

Comments: 1. The plot clearly depicts a robust negative relationship between mental distress or disorder and life satisfaction. As indicated by the negative slope, an increase in mental distress or disorder corresponds to a decrease in life satisfaction. 2. The negative slope emphasizes that escalating levels of mental distress or disorder are associated with reduced life satisfaction.

```
# Plot relationship between everyday discrimination and 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))
```



Comments: 1. The plot clearly demonstrates a negative relationship between everyday discrimination and life satisfaction. 2. The negative slope indicates that as everyday discrimination increases, life satisfaction tends to decrease.

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.

Solution: Justification for choosing the predictor as Mental Distress/Disorder: 1. The plot for Mental Distress/Disorder indicates a strong negative relationship between mental distress and life satisfaction. 2. This suggests that as mental distress increases, life satisfaction decreases. 3. The plot also demonstrates a strong correlation, as evidenced by the clustering of data points around the fitted line of the regression model. 4. Given these findings, I decided to include Mental Distress/Disorder as a predictor in the model, despite other predictors also displaying strong negative relationships.

Fit the model

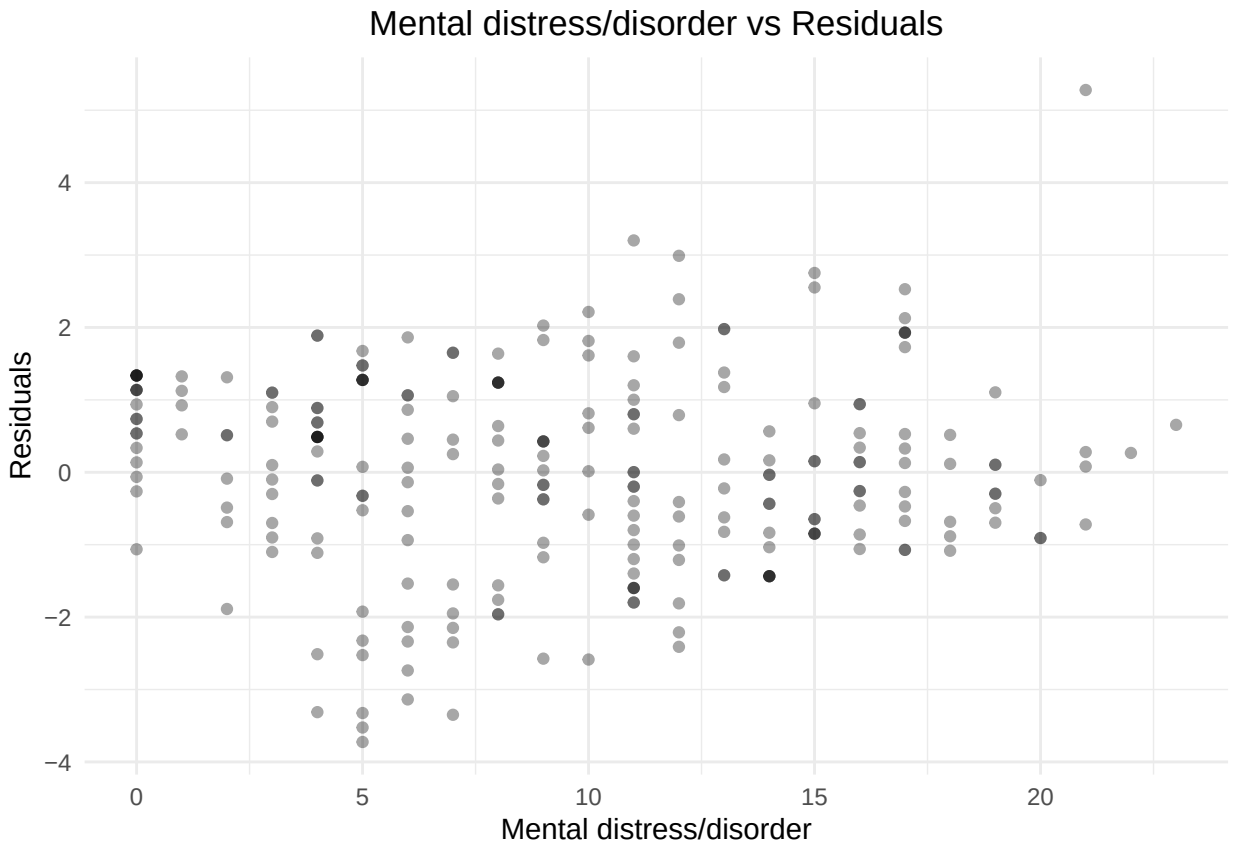
```
# Fit a linear regression model to assess the relationship between LIFESAT_I (life satisfaction) and KE
model1 <- lm(LIFESAT_I ~ KESSLER6_I, data = my_data)
summary(model1)
```

```
##
```

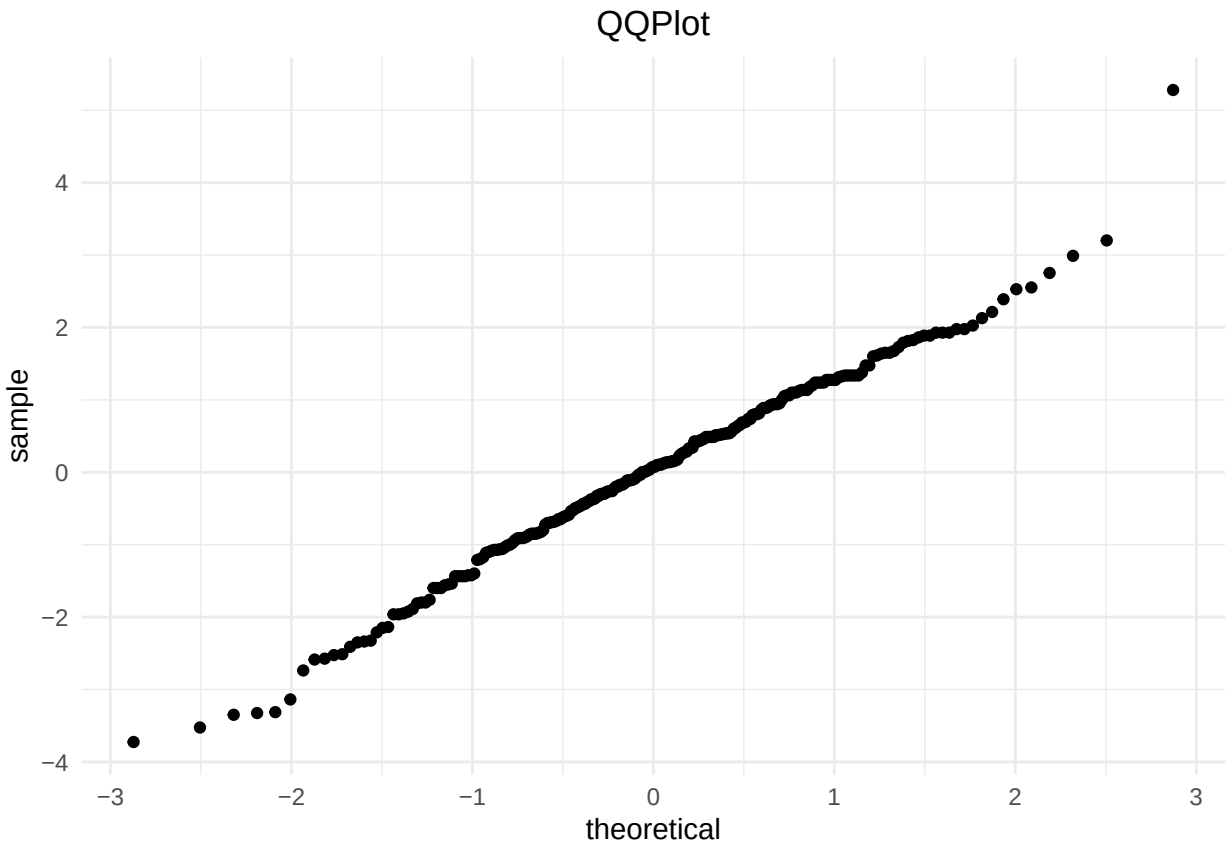
```
## Call:
## lm(formula = LIFESAT_I ~ KESSLER6_I, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7246 -0.8471  0.0754  0.9407  5.2794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.66336    0.16718   33.88   <2e-16 ***
## 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)
# Plotting residuals against mental distress/disorder
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))
```




```
# Add residuals to the dataset from model1
my_data %>%
  add_residuals(model1, "resid") %>%
  # Create a ggplot with residuals for a QQ plot
  ggplot(aes(sample=resid)) +
  # Add a QQ plot using geom_qq()
  geom_qq() +
  # Set the title of the plot
  ggtitle("QQPlot") +
  # Apply a minimal theme to the plot
  theme_minimal() +
  # Center the plot title
  theme(plot.title = element_text(hjust = 0.5))
```



Comments on Residual Plots and Outliers: 1. The QQ-Plot reveals the presence of outliers, which can significantly impact the accuracy of model predictions. 2. Outliers should be addressed to improve the model's performance. 3. It's important to remove outliers to enhance the reliability of the model. 4. Upon examining the residual plot, it appears to display simple random scatter, indicating no violations of model assumptions. 5. However, one data point with a residual greater than 4 stands out, suggesting it could be an outlier. 6. Outliers have the potential to influence the slope and intercept of the fitted line, leading to less precise predictions. 7. Consequently, the outlier has been removed from the data, and the model has been refitted to reassess model diagnostics.

Re-fit the model after removing outlier

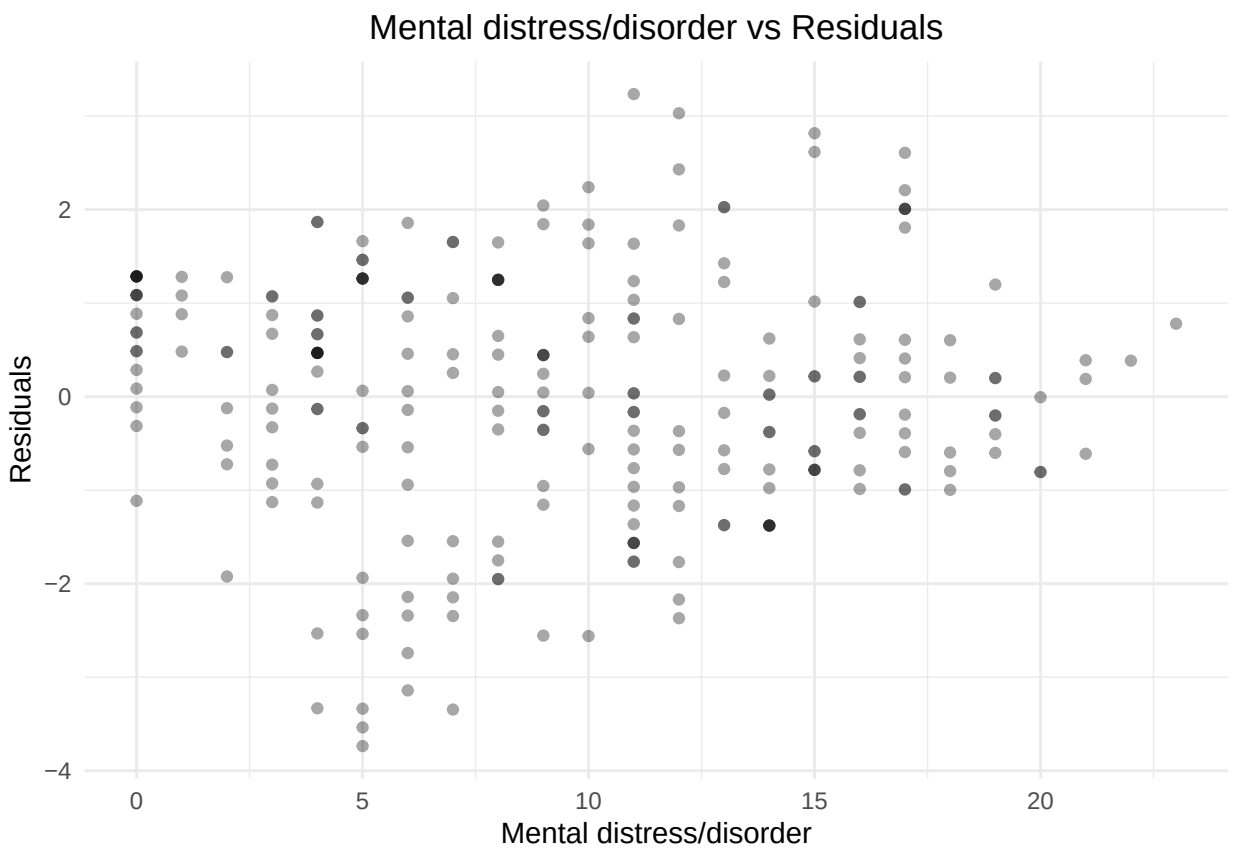
```r
# Add residuals to the dataset from model1 and filter out residuals less than or equal to 4
my_data <- my_data %>%
  add_residuals(model1, "resid") %>%
  filter(resid <= 4)

# Fit a new linear regression model using lm() function with filtered data
new_model1 <- lm(LIFESAT_I ~ KESSLER6_I, data = my_data)

# Summary of the new linear regression model
summary(new_model1)
```

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

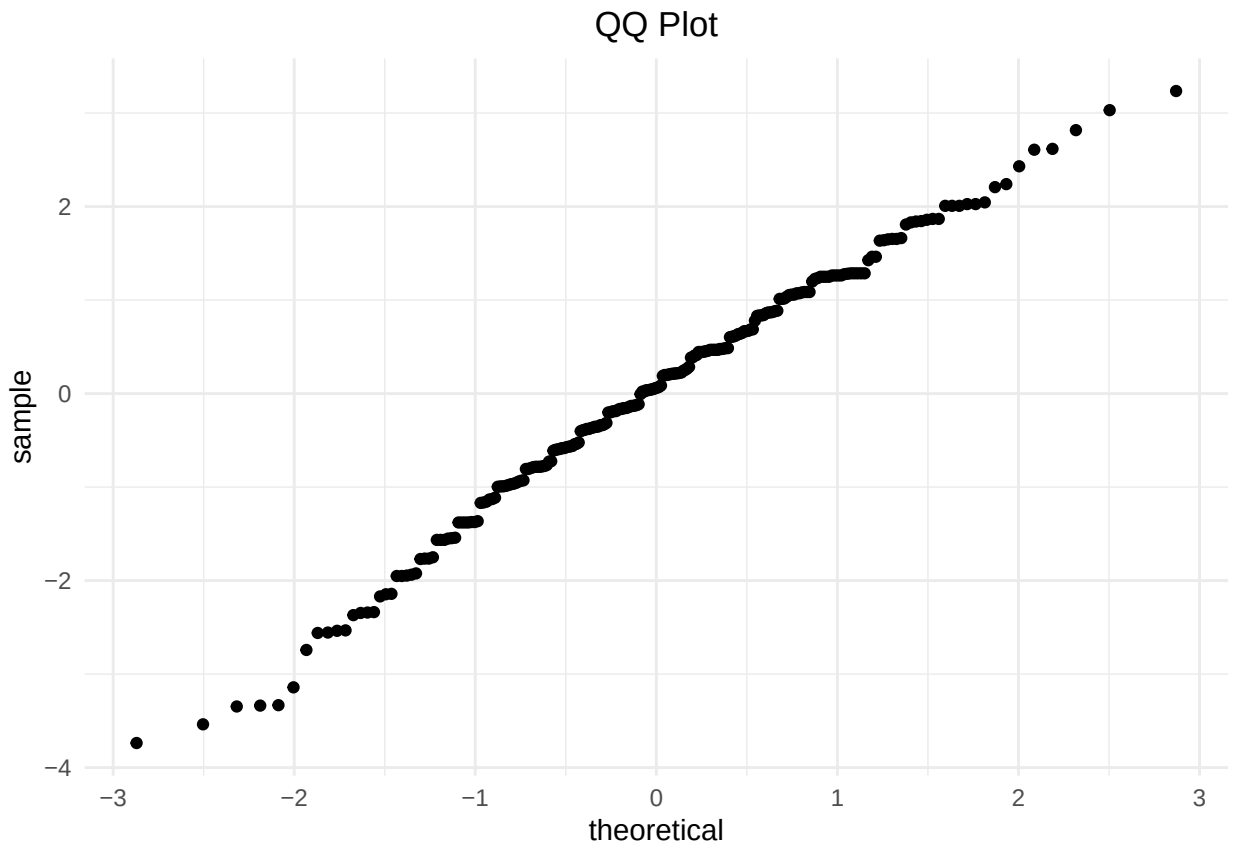
```
##     Min      1Q  Median      3Q     Max
## -3.7369 -0.7841  0.0608  0.9177  3.2354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.71386    0.16259   35.14   <2e-16 ***
## KESSLER6_I  -0.19539    0.01473  -13.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 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

```
# Add residuals from the new linear regression model (new_model1) to the dataset my_data
# Then, create a scatter plot to visualize the relationship between KESSLER6_I (mental distress/disorde
my_data %>%
  add_residuals(new_model1, "resid") %>%  # Adding residuals from new_model1 to my_data
  ggplot(aes(x = KESSLER6_I)) +  # Setting x-axis to KESSLER6_I
  geom_point(aes(y = resid), alpha = 0.35) +  # Adding scatter plot with residuals on y-axis
  labs(x = "Mental distress/disorder", y = "Residuals") +  # Adding labels to axes
  ggtitle("Mental distress/disorder vs Residuals") +  # Adding title to the plot
  theme_minimal() +  # Applying minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centering the title
```



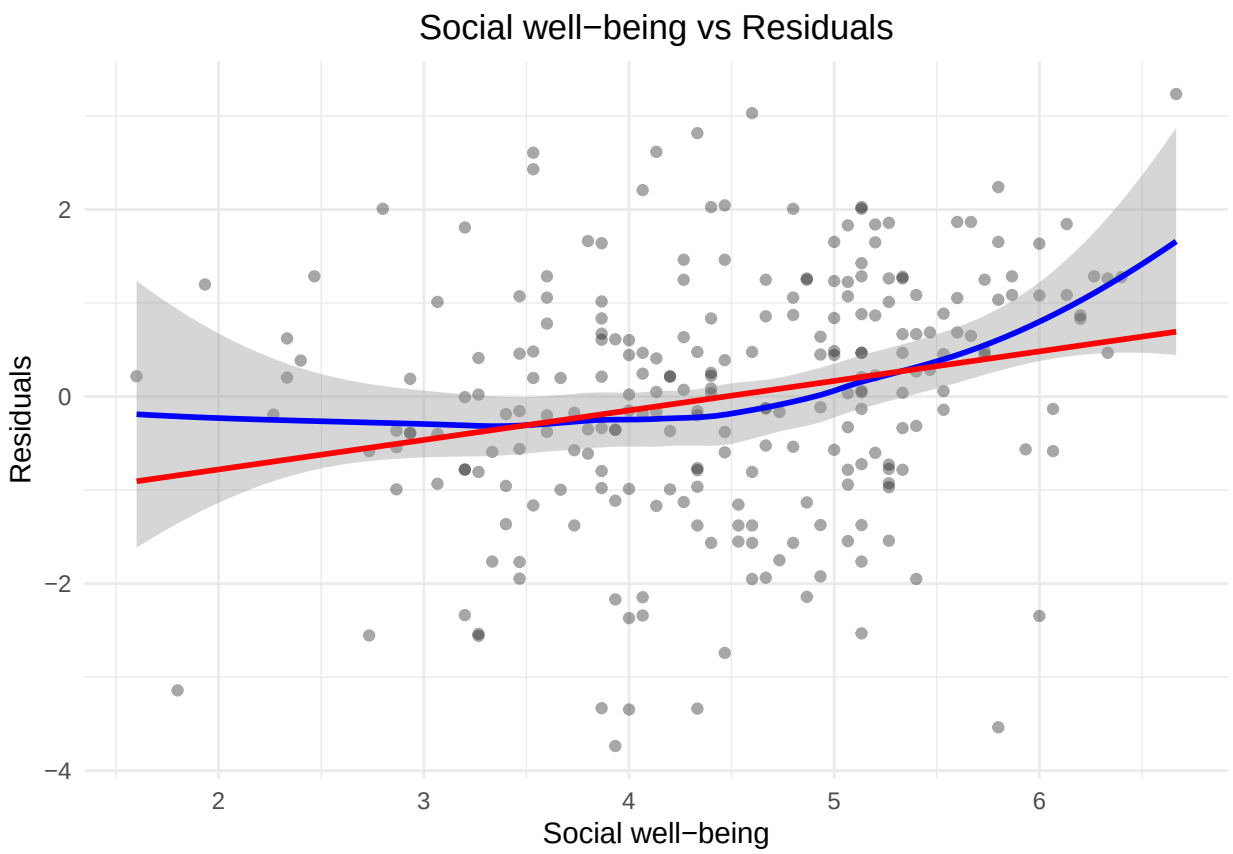
```r
# Add residuals to my_data using new_model1 and name the new column "resid"
my_data %>%
  add_residuals(new_model1, "resid") %>%
  # Create a QQ plot of the residuals
  ggplot(aes(sample=resid)) +
  geom_qq() +  # Add QQ plot
  ggtitle("QQ Plot") +  # Set plot title
  theme_minimal() +  # Set minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Center plot title
```



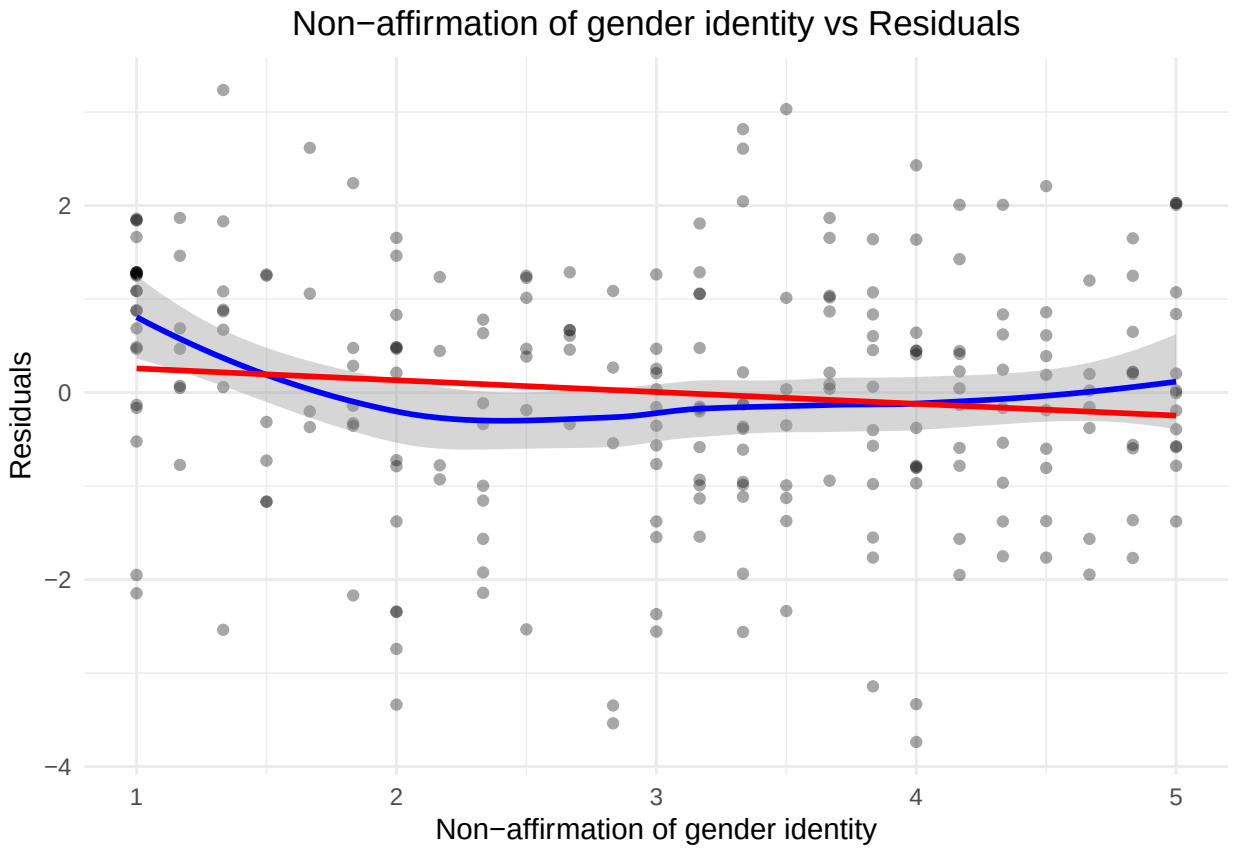
Comments: It can be noted that there are no outliers present at this point, indicating that the current model performs better than the previous one that contained outliers.

Problem 5

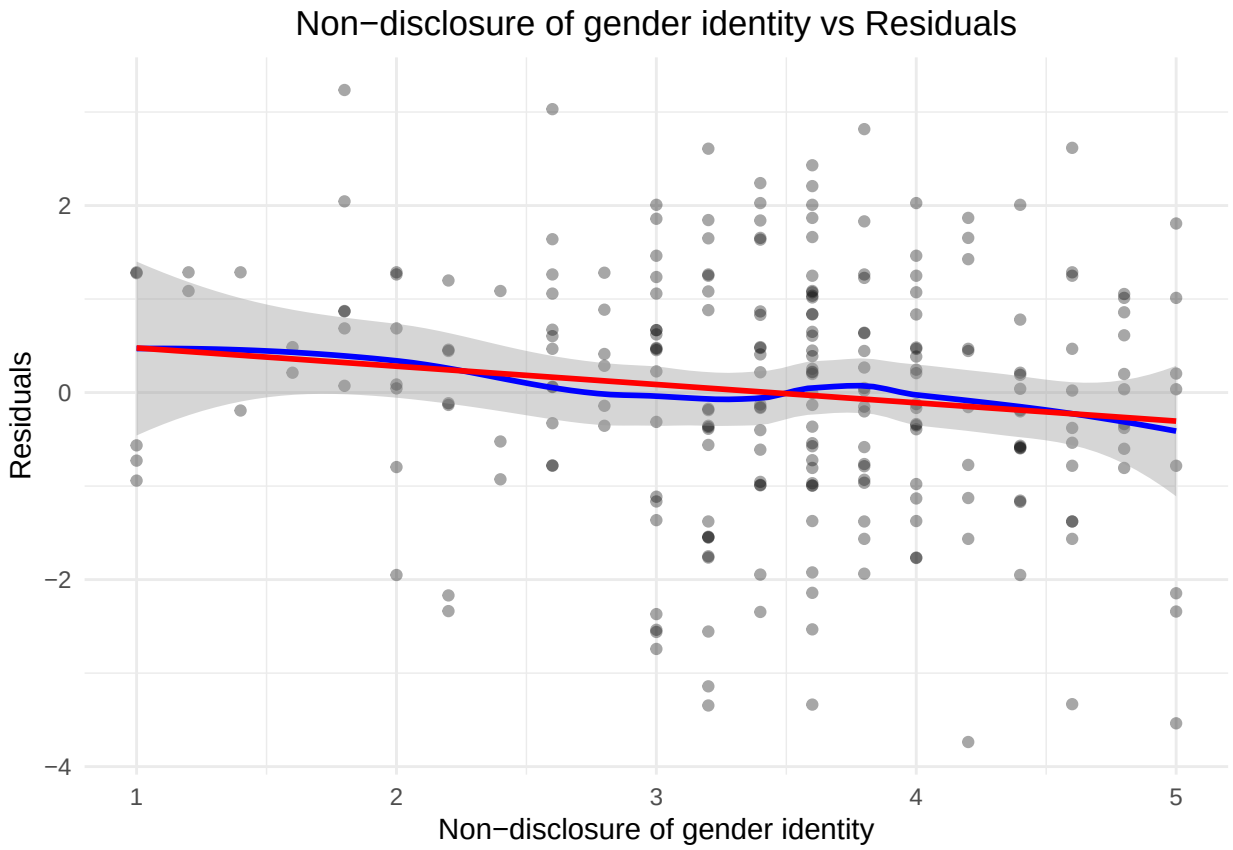
```r
# Plotting the relationship between Social well-being (SOCIALWB_I) and Residuals
my_data %>%
  # Adding residuals from the new_model1 to the dataset
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = SOCIALWB_I)) +  # X-axis: Social well-being
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +  # Loess smoothing
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +  # Linear
  labs(x = "Social well-being", y = "Residuals") +  # Labels for axes
  ggtitle("Social well-being vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```



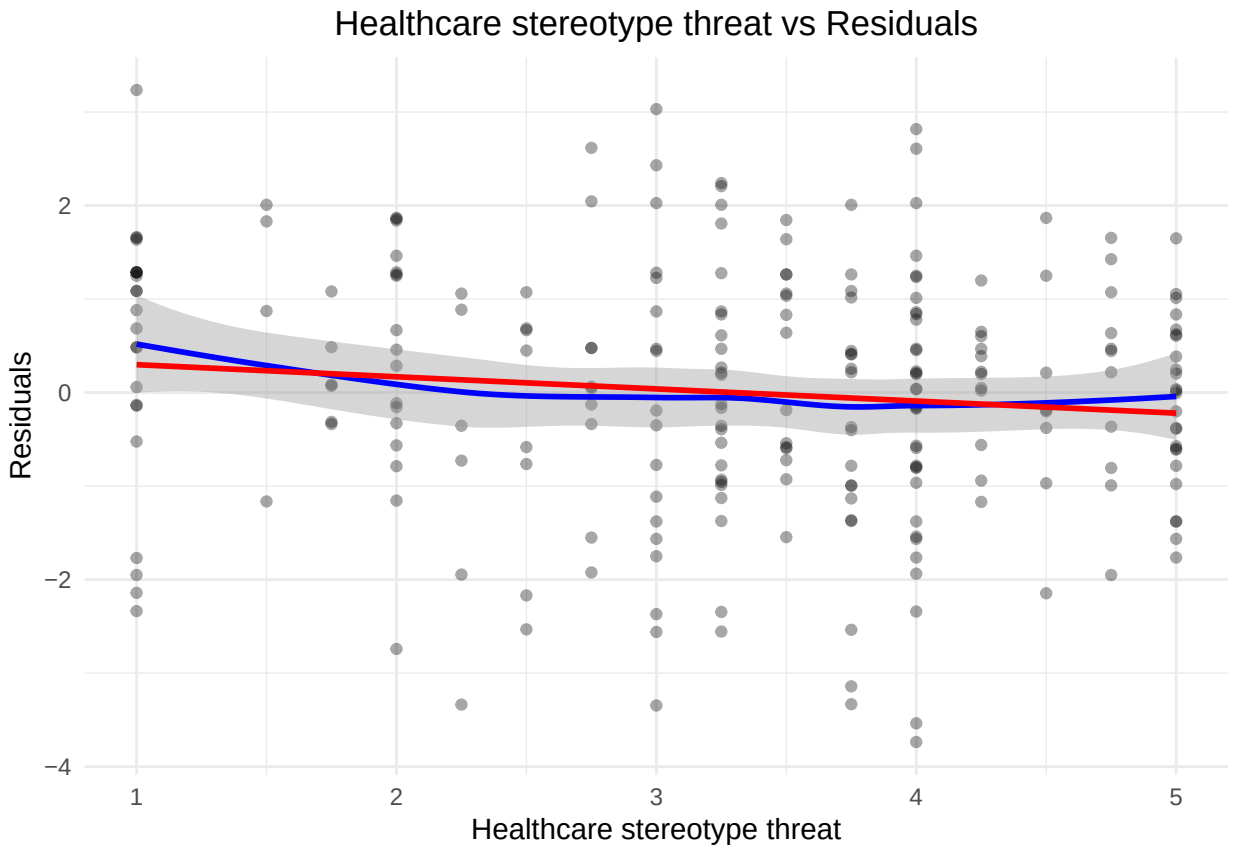

```
# Plotting the relationship between Non-affirmation of gender identity (NONAFFIRM_I) and Residuals
my_data %>%
  # Adding residuals from the new_model1 to the dataset
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = NONAFFIRM_I)) +  # X-axis: Non-affirmation of gender identity
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +  # Loess smoothing
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +  # Linear
  labs(x = "Non-affirmation of gender identity", y = "Residuals") +  # Labels for axes
  ggtitle("Non-affirmation of gender identity vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```



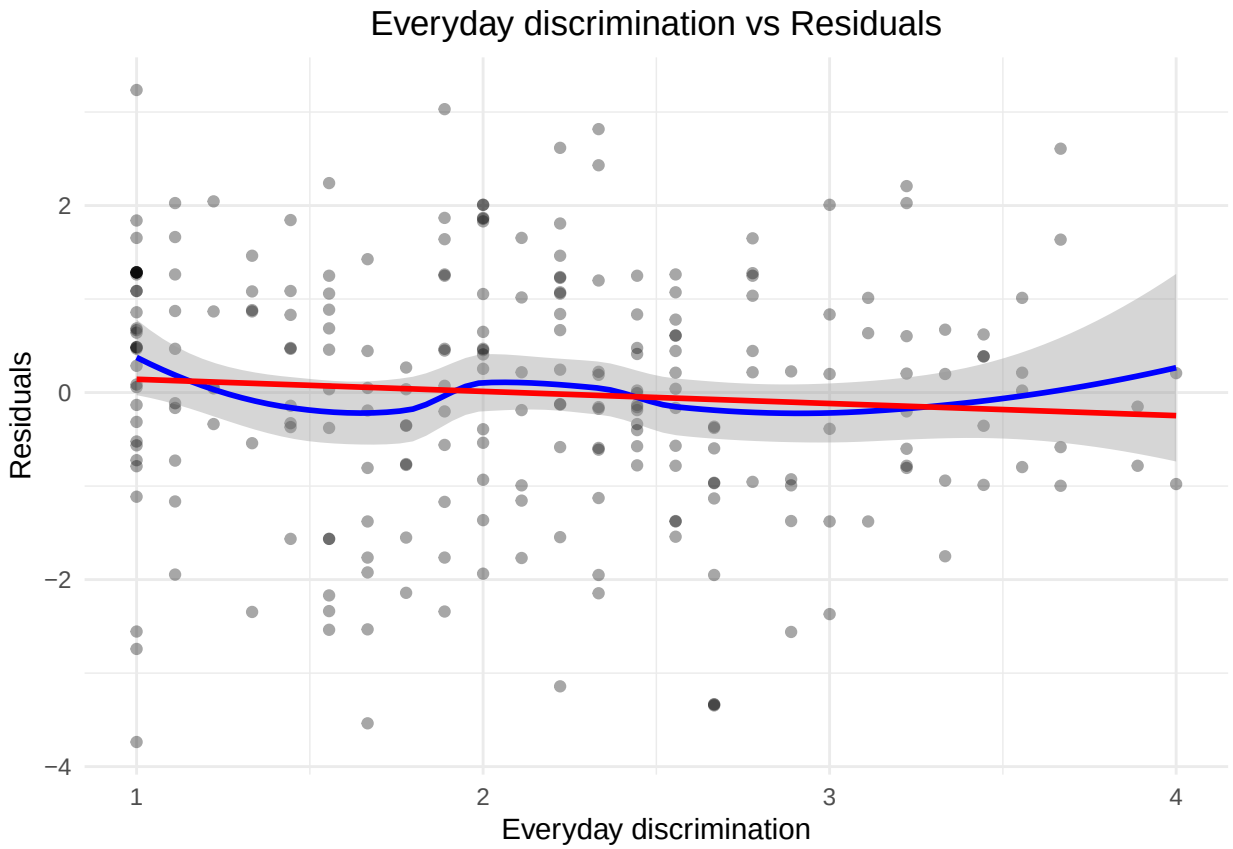

```r
# Plotting the relationship between Non-disclosure of gender identity (NONDISCLOSURE_I) and Residuals
my_data %>%
  # Adding residuals from the new_model1 to the dataset
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = NONDISCLOSURE_I)) +  # X-axis: Non-disclosure of gender identity
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +  # Loess smoothing
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +  # Linear
  labs(x = "Non-disclosure of gender identity", y = "Residuals") +  # Labels for axes
  ggtitle("Non-disclosure of gender identity vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```




```
# Plotting the relationship between Healthcare stereotype threat (HCTHREAT_I) and Residuals
my_data %>%
  # Adding residuals from the new_model1 to the dataset
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = HCTHREAT_I)) +  # X-axis: Healthcare stereotype threat
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +  # Loess smoothing
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +  # Linear
  labs(x = "Healthcare stereotype threat", y = "Residuals") +  # Labels for axes
  ggtitle("Healthcare stereotype threat vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```




```r
# Plotting the relationship between Everyday discrimination (EVERYDAY_I) and Residuals
my_data %>%
  # Adding residuals from the new_model1 to the dataset
  add_residuals(new_model1, "resid") %>%
  ggplot(aes(x = EVERYDAY_I)) +  # X-axis: Everyday discrimination
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  geom_smooth(aes(y = resid), color = "blue", method = 'loess', formula = 'y ~ x') +  # Loess smoothing
  geom_smooth(aes(y = resid), method = "lm", se = FALSE, formula = 'y ~ x', color = "red") +  # Linear
  labs(x = "Everyday discrimination", y = "Residuals") +  # Labels for axes
  ggtitle("Everyday discrimination vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```



After analyzing the plots, I've decided to include Social well-being (SOCIALWB_I) along with KESSLER6_I (Mental distress/disorder) in my model. The reason behind this decision is that I observed a stronger positive trend and a closer alignment with the linear line in the plot for SOCIALWB_I. This suggests that SOCIALWB_I has a significant positive correlation with life satisfaction (LIFESAT_I) and can offer valuable insights into the analysis. By incorporating predictors with diverse relationships to the outcome, such as the positive association with SOCIALWB_I and the negative one with KESSLER6_I, I aim to gain a deeper understanding of the complex factors impacting life satisfaction. This approach takes into consideration both the adverse effects of mental distress and the beneficial contributions of social well-being.

Fit the model

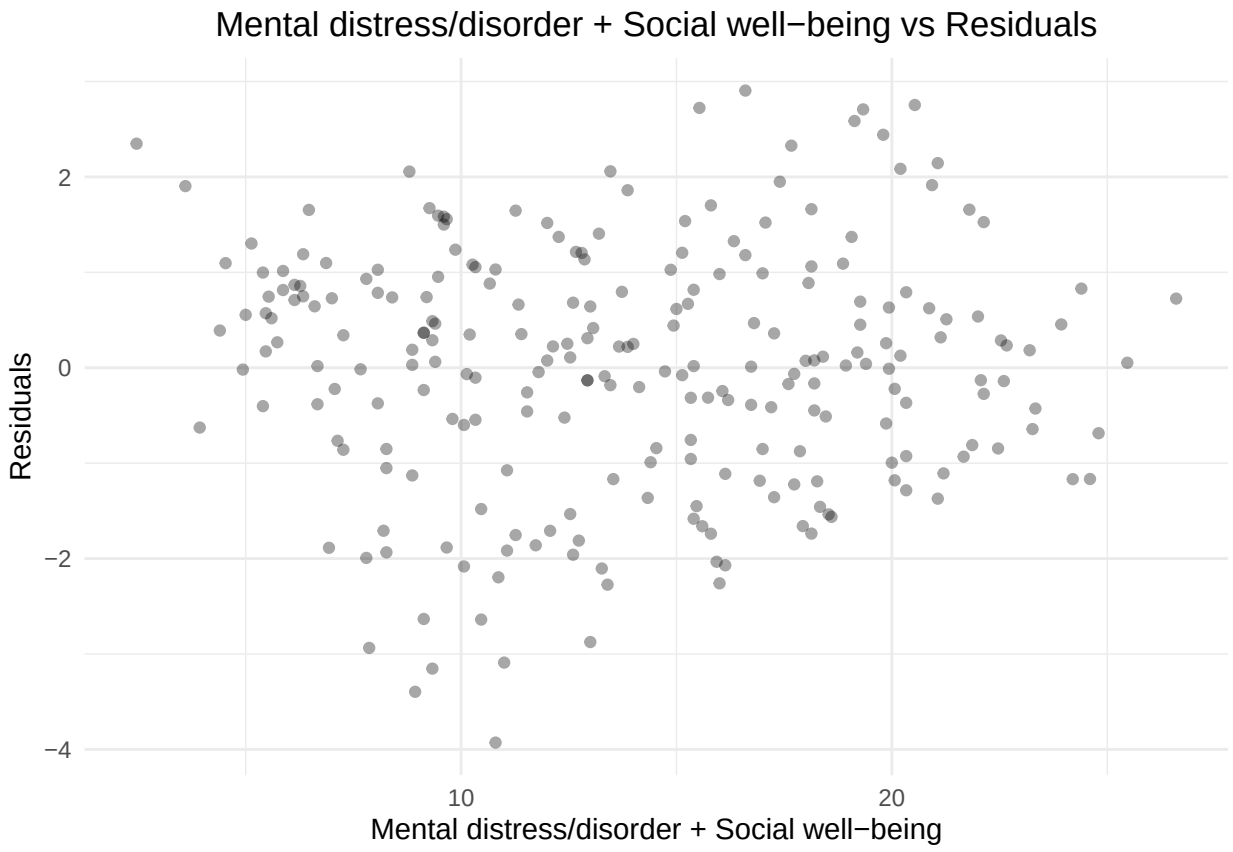
```
# Fit a linear regression model using lm() function
model2 <- lm(LIFESAT_I ~ KESSLER6_I + SOCIALWB_I, data = my_data)

# Summary of the linear regression model
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

```
# Plotting the relationship between the combination of Mental distress/disorder (KESSLER6_I) and Social
my_data %>%
  # Adding residuals from model2 to the dataset
  add_residuals(model2, "resid") %>%
  ggplot(aes(x = KESSLER6_I + SOCIALWB_I)) +  # X-axis: Combination of Mental distress/disorder and Soc
  geom_point(aes(y = resid), alpha = 0.35) +  # Scatter plot of residuals
  labs(x = "Mental distress/disorder + Social well-being", y = "Residuals") +  # Labels for axes
  ggtitle("Mental distress/disorder + Social well-being vs Residuals") +  # Title for the plot
  theme_minimal() +  # Minimal theme
  theme(plot.title = element_text(hjust = 0.5))  # Centered title
```




```
# Creating a QQ plot to assess the normality of residuals from model2
my_data %>%
  # Adding residuals from model2 to the dataset
  add_residuals(model2, "resid") %>%
```

```r
ggplot(aes(sample=resid)) +  # Q-Q plot with residuals as sample
geom_qq() +  # Adding the quantile-quantile plot
ggtitle("QQPlot") +  # Title for the plot
theme_minimal() +  # Applying a minimal theme
theme(plot.title = element_text(hjust = 0.5))  # Centering the title
```